Welcome

Co-clustering Analysis of Multidimensional Big Data

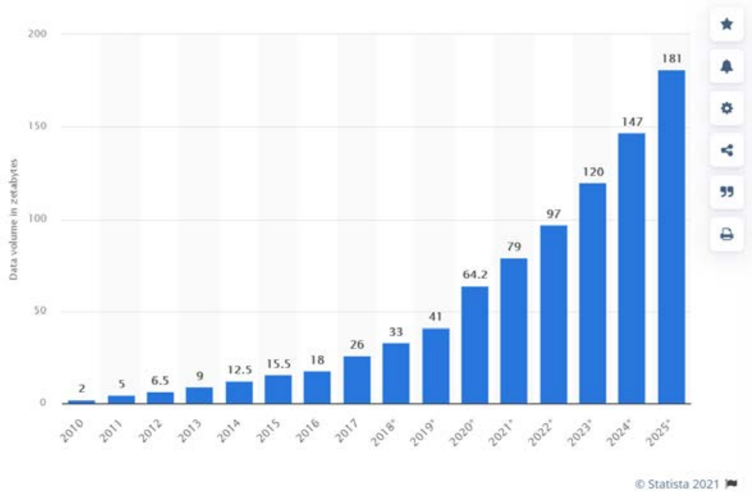
Hong Yan
City University of Hong Kong
Centre for Intelligent Multidimensional Data Analysis Limited

Group Members

People who have contributed to this work:

- Lijiang Chen
- Long Chen
- Ray Cheung
- Dao-Qing Dai
- Xiangchao Gan
- Lixin Han
- Sheheryar Khan
- Lichun Ma
- Alan W C Liew
- Benben Liu
- Xinyu Liu
- David Smith
- Peter Tino
- Debby Wang
- Doris Z Wang
- Wen-Hui Yang
- Chiwai Yu
- Hongya Zhao
- Xuefei Zhe
- Bin Zou

Multidimensional Big Data



Reference: https://www.statista.com/statistics/871513/worldwide-data-created/

$$\text{Data}[x_1][x_2]\cdots[x_N]$$

$$\mathcal{A} = \left(a_{i_1 i_2 \cdots i_N}\right) \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_N}$$

Higher-order tensor

Multi-way data array

Multidimensional data

Example 1: Gene × Condition × Time

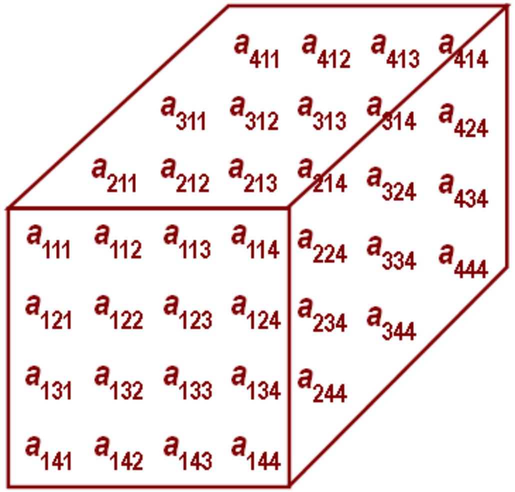
Example 2: Document × Key word × Region

Example 3: Speech × Frequency × Time

0, 1, ..., 9,
½, e, π,
a, b, c, x, y, ...

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix}$$

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$



Number → **Vector** → **Matrix** $\xrightarrow{???}$ **Tensor**

Numerical Computation and Data Analysis

Numerical Computation:

FORTRAN: 1950’s, SVD: 1950’s~70’s

LINPACK, EISPACK, MINPACK: 1970’s~80’s

MATLAB: 1980’s, R: 1990’s, LAPACK: 1990’s

Information Processing:

Digital signal, image and video processing

Data acquisition, storage, retrieval, transmission, security

Computational linguistics, internet search, business intelligence

Artificial Intelligence (Data Driven):

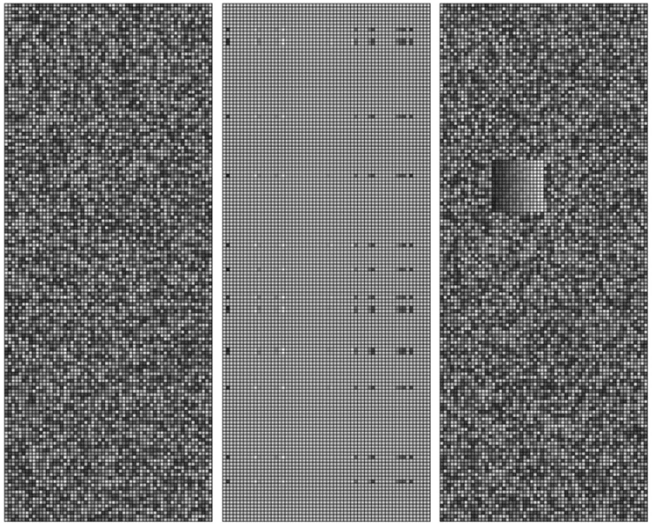
Hot

Machine learning, deep learning

Intelligent robot and autonomous vehicles

Data mining, pattern matching and decision making

Focus of This Talk: Coherent Pattern Detection



Coherent Pattern Detection:

- Input: big multidimensional data
- Output: smaller coherent patterns

Techniques:

- Matrix / tensor decomposition
- Low-rank matrix / tensor identification
- Hyperplane models for coherent pattern detection

Applications:

- Disease diagnosis based on gene expression data
- Cell division data analysis
- Removal of irrelevant features
- Protein secondary structure prediction
- Human facial expression analysis and classification

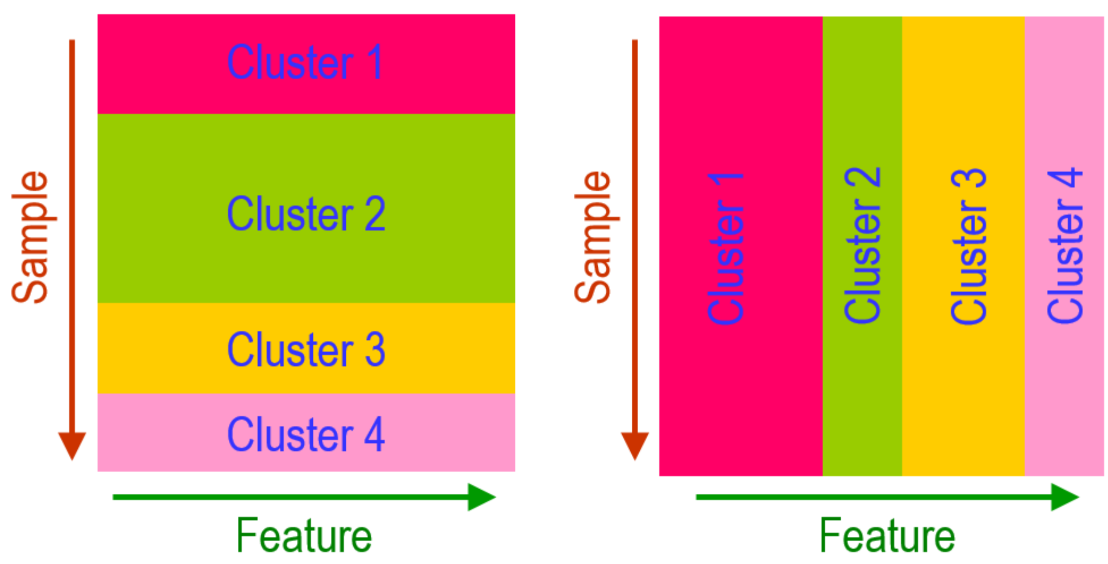
Stigma:

* NP-hard
* Inherently intractable

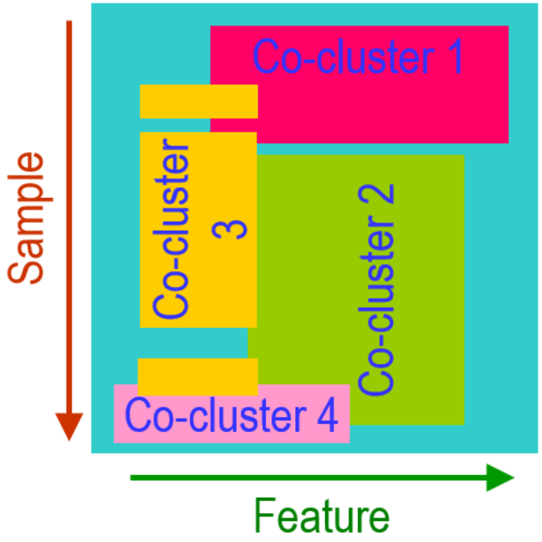
Machine Learning Methods

- Machine Learning
  - Supervised Learning
    - Bayes
    - Decision region formation:
      - Lines, curves, planes, spheres
    - Decision trees, Random forests
    - Linear / nonlinear / logistic regression
    - Support vector machines
    - Artificial neural networks
  - Unsupervised Learning
    - Dimensionality reduction (PCA, SVD, …)
    - Hidden Markov model
    - Clustering, k-means, hierarchical, …
    - Graph cutting
    - Minimum spanning trees
    - Hough transform
    - Self-organizing maps

Clustering and Co-clustering



Classification in one direction only



Classification in both directions

Differences between PCA and Co-clustering

$\mathbf{x}_1 = (x_{11}, x_{12}, \ldots, x_{1n})$

$\mathbf{x}_2 = (x_{21}, x_{22}, \ldots, x_{2n})$

……

$\mathbf{x}_m = (x_{m1}, x_{m2}, \ldots, x_{mn})$

Input data:

$m \times n$ matrix

m samples

n features

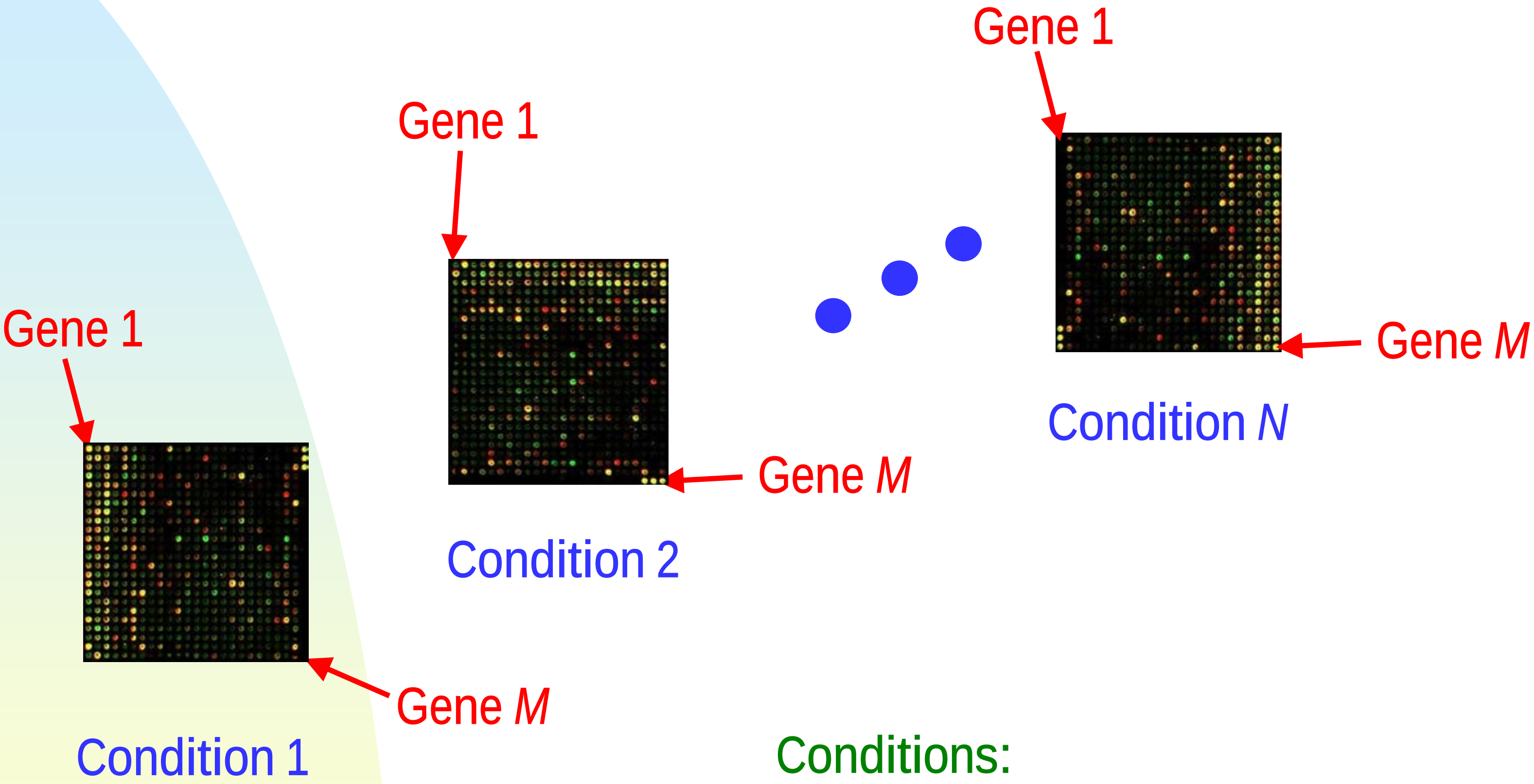
PCA: compute PCs

All samples and all features contribute to the PCs

Co-clustering: look for subsets

Some samples and some features contribute to the co-clusters

Genes and Conditions



Conditions:

Time points, organs, species, etc

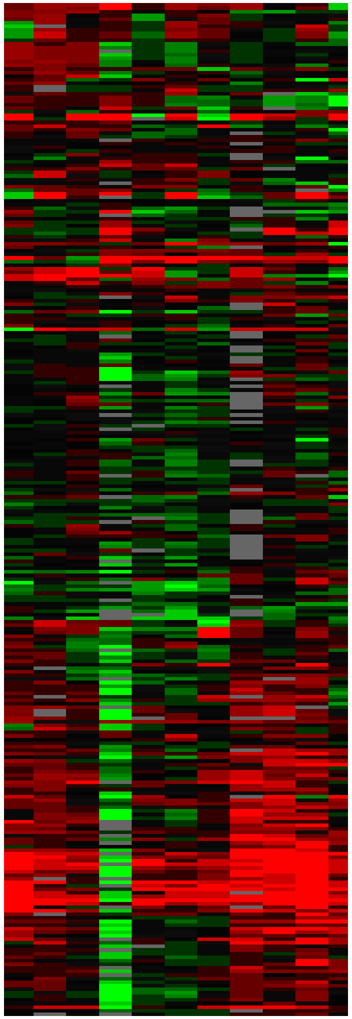
Microarray Data Matrix

N Conditions

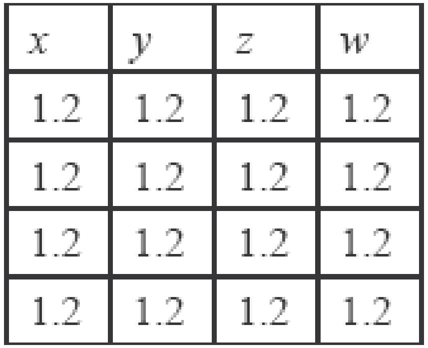
$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{M1} & x_{M2} & \cdots & x_{MN} \end{bmatrix}$$

M Genes

$$x_{ij} = \log \frac{\overline{R}_{ij}^{\text{fearure}} - \overline{R}_{ij}^{\text{background}}}{\overline{G}_{ij}^{\text{fearure}} - \overline{G}_{ij}^{\text{background}}}$$

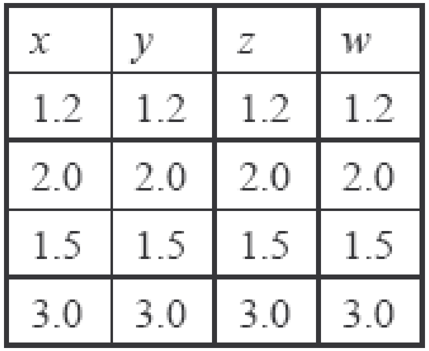


Different Types of Co-clusters



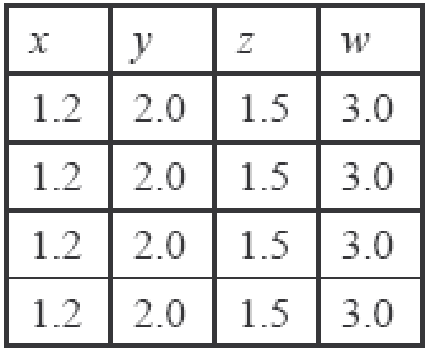
x	*y*	*z*	*w*
1.2	1.2	1.2	1.2
1.2	1.2	1.2	1.2
1.2	1.2	1.2	1.2
1.2	1.2	1.2	1.2

Constant



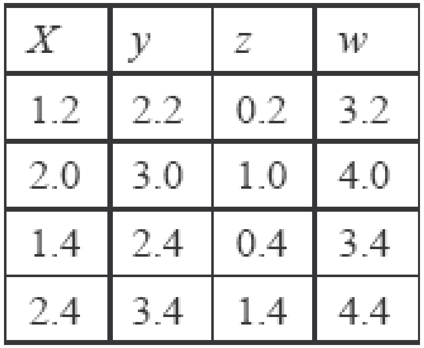
x	*y*	*z*	*w*
1.2	1.2	1.2	1.2
2.0	2.0	2.0	2.0
1.5	1.5	1.5	1.5
3.0	3.0	3.0	3.0

Constant row



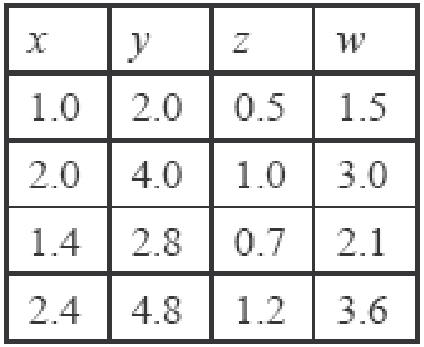
x	*y*	*z*	*w*
1.2	2.0	1.5	3.0
1.2	2.0	1.5	3.0
1.2	2.0	1.5	3.0
1.2	2.0	1.5	3.0

Constant column



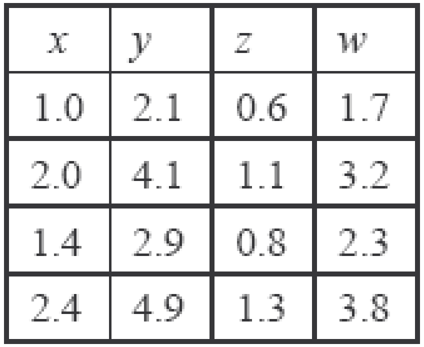
X	*y*	*z*	*w*
1.2	2.2	0.2	3.2
2.0	3.0	1.0	4.0
1.4	2.4	0.4	3.4
2.4	3.4	1.4	4.4

Additive



x	*y*	*z*	*w*
1.0	2.0	0.5	1.5
2.0	4.0	1.0	3.0
1.4	2.8	0.7	2.1
2.4	4.8	1.2	3.6

Multiplicative



x	*y*	*z*	*w*
1.0	2.1	0.6	1.7
2.0	4.1	1.1	3.2
1.4	2.9	0.8	2.3
2.4	4.9	1.3	3.8

Linear

Example of Existing Methods

Cheng and Church’s Algorithm

$$H(I,J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} \left(e_{ij} - e_{iJ} - e_{Ij} + e_{IJ} \right)^2$$

$$e_{Ij} = \frac{1}{|I|} \sum_{i \in I} e_{ij}$$

Column average

I: subset of genes
J: subset of conditions
I X ***J***: bicluster

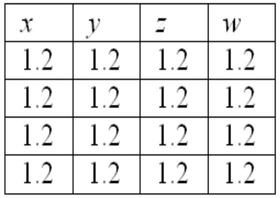
$$e_{iJ} = \frac{1}{|J|} \sum_{j \in J} e_{ij}$$

Row average

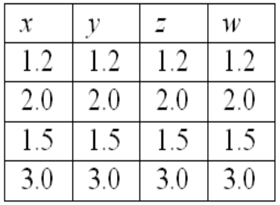
$$e_{IJ} = \frac{1}{|I||J|} \sum_{i \in I, j \in J} e_{ij}, e_{ij}$$

Pattern average

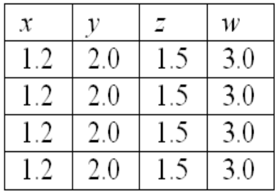
Geometrical Interpretations: Our Approach



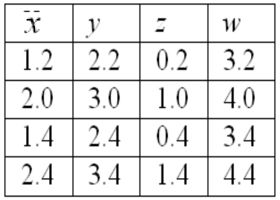
x	y	z	w
1.2	1.2	1.2	1.2
1.2	1.2	1.2	1.2
1.2	1.2	1.2	1.2
1.2	1.2	1.2	1.2



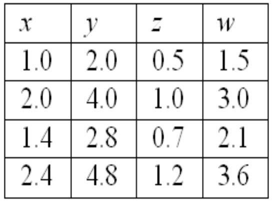
x	y	z	w
1.2	1.2	1.2	1.2
2.0	2.0	2.0	2.0
1.5	1.5	1.5	1.5
3.0	3.0	3.0	3.0



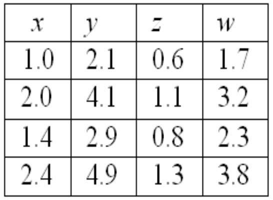
x	y	z	w
1.2	2.0	1.5	3.0
1.2	2.0	1.5	3.0
1.2	2.0	1.5	3.0
1.2	2.0	1.5	3.0



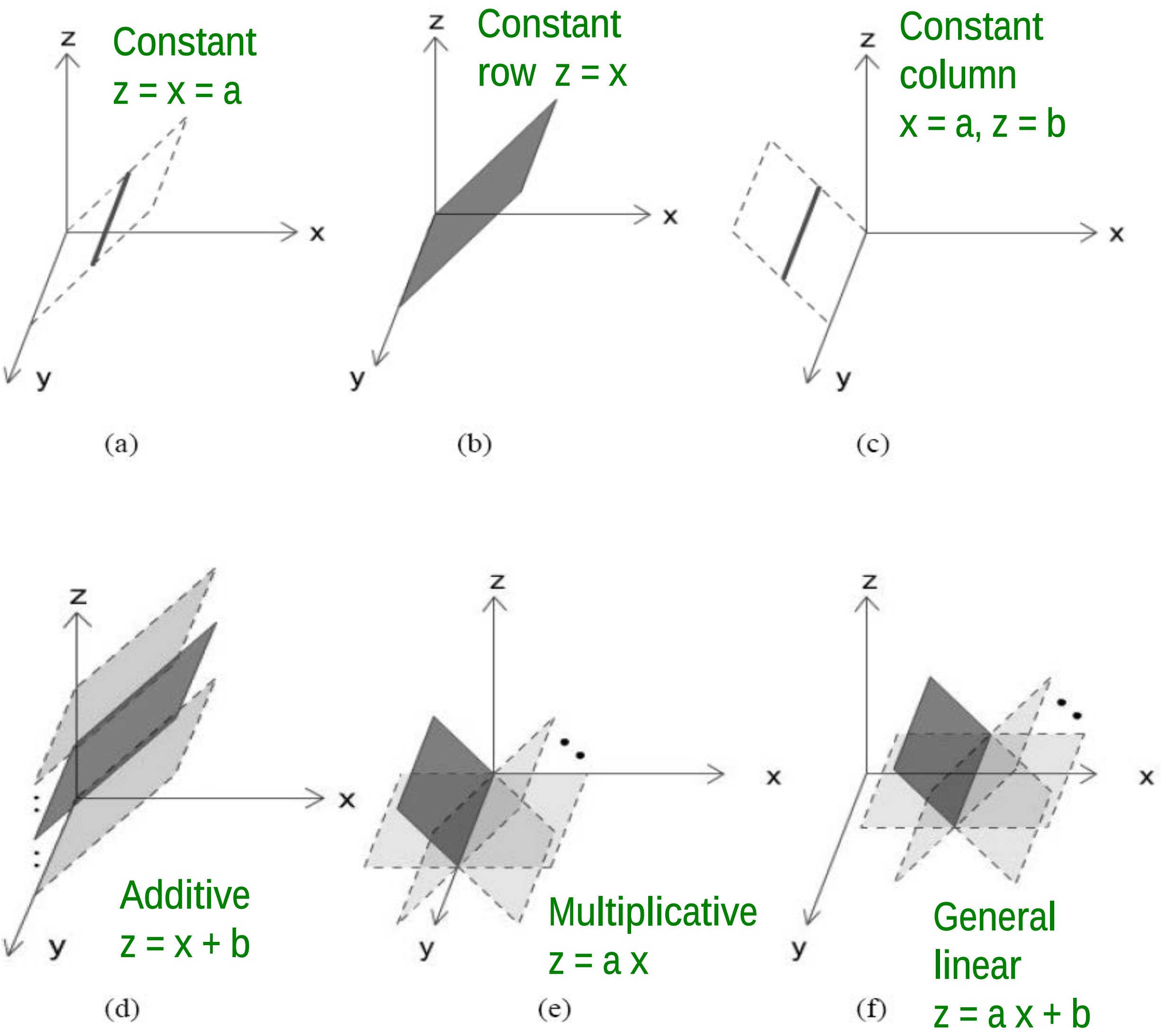
x	y	z	w
1.2	2.2	0.2	3.2
2.0	3.0	1.0	4.0
1.4	2.4	0.4	3.4
2.4	3.4	1.4	4.4



x	y	z	w
1.0	2.0	0.5	1.5
2.0	4.0	1.0	3.0
1.4	2.8	0.7	2.1
2.4	4.8	1.2	3.6



x	y	z	w
1.0	2.1	0.6	1.7
2.0	4.1	1.1	3.2
1.4	2.9	0.8	2.3
2.4	4.9	1.3	3.8



(a) (b) (c)

(d) (e) (f)

Constant
z = x = a

Constant
row z = x

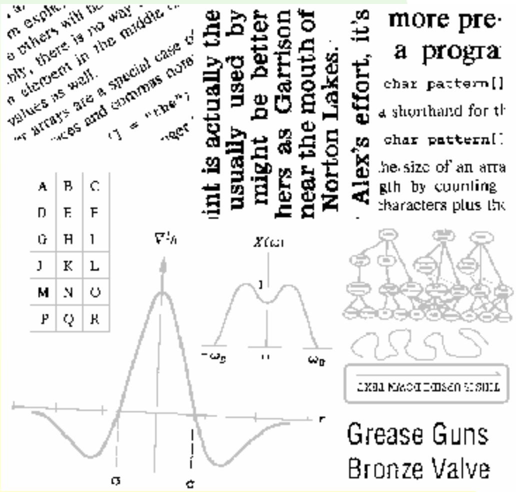
Constant
column
x = a, z = b

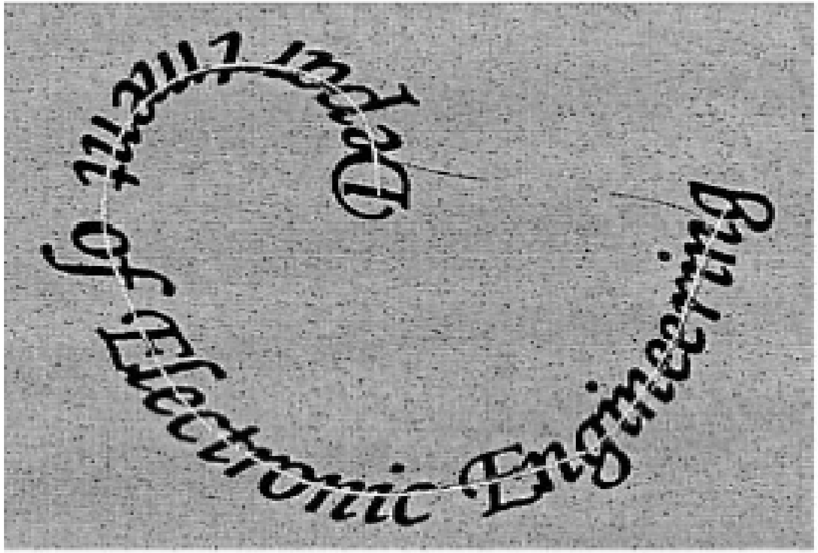
Additive
z = x + b

Multiplicative
z = a x

General
linear
z = a x + b

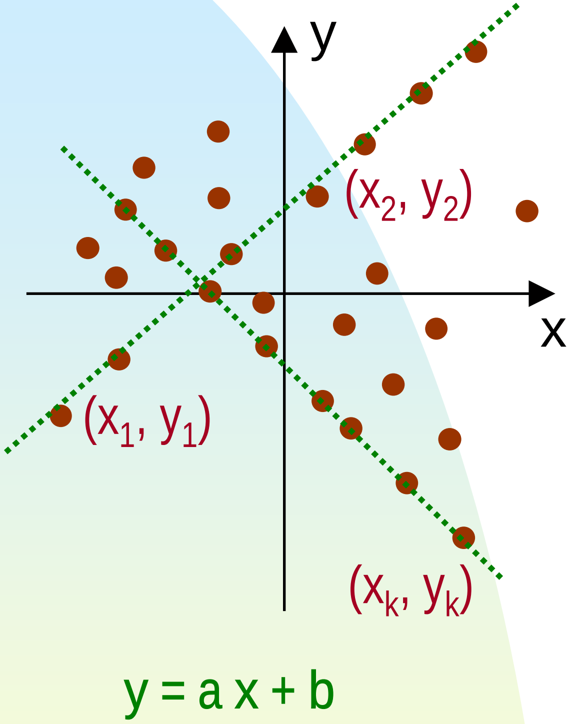
The Hough Transform



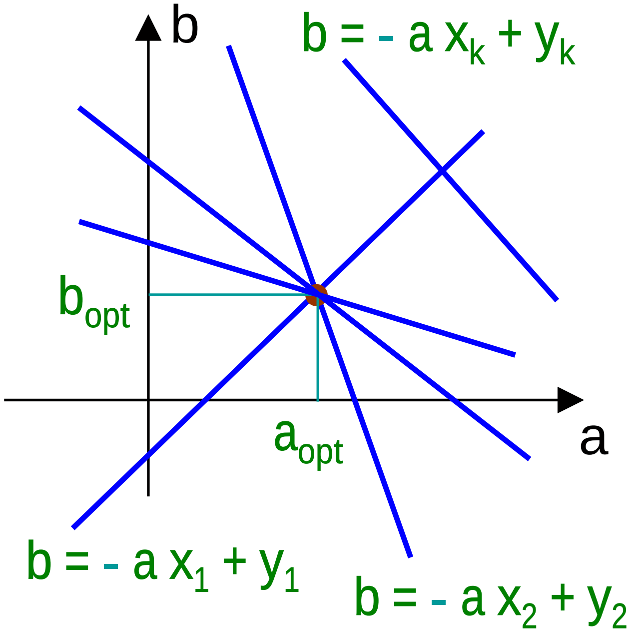


References: H Yan, *IEEE Trans. SMC, Part B*, 34(1):210-221, 2004; H Yan, "Curve tracing system," *United States Patent* 7263538, 2007.

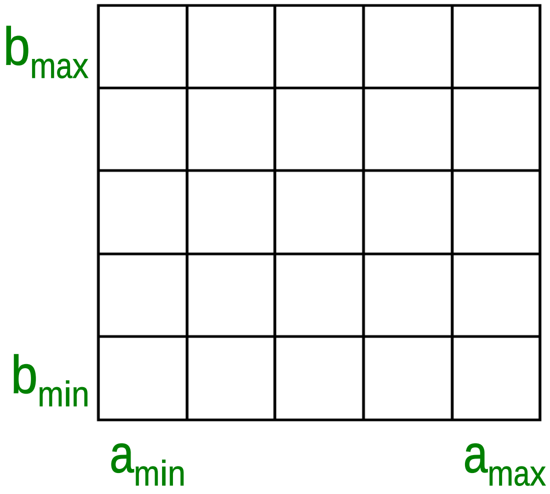
Lines in the Hough Space



Data space



Hough space



Quantized parameters
(Voting ballot counting)

Similar Method: Radon Transform (projections at different angles)

Hyperplane in Object Space

$$F_0(j) = \sum_{i=1}^{N-1} \beta_i F_i(j) + \beta_N$$ Hyperplane equation

$$F_0(j),\ F_1(j),\ \ldots,\ F_{N-1}\ (j),\ \ (j = 1, 2, \ldots, M)$$ Condition variables

$$\beta_1, \beta_2, \ \ldots, \beta_N$$ Hyperplane coefficients

$$\sum_{i=1}^{M-1} F_i(j)\beta_i + \beta_M - F_0(j) = 0 \qquad 0 \le j \le M$$

$$\beta_i \in [C_i - L_i, C_i + L_i]$$ Parameter value range

Hyperplane in Parameter Space

$$\sum_{i=1}^{N-1} \frac{F_i(j)L_i}{W(j)L_N} \frac{\beta_i}{L_i} + \frac{\beta_N}{W(j)L_N} - \frac{F_0(j)}{W(j)L_N} = 0$$

Original hyperplane equation in object space

$$\frac{F_i(j)L_i}{W(j)L_N} = a_i(j)$$

Hyperplane coefficients

$$X_i = \frac{\beta_i}{L_i}$$

Hyperplane variables

$$\sum_{i=1}^{N} a_i^2(j) = 1$$

Normalization

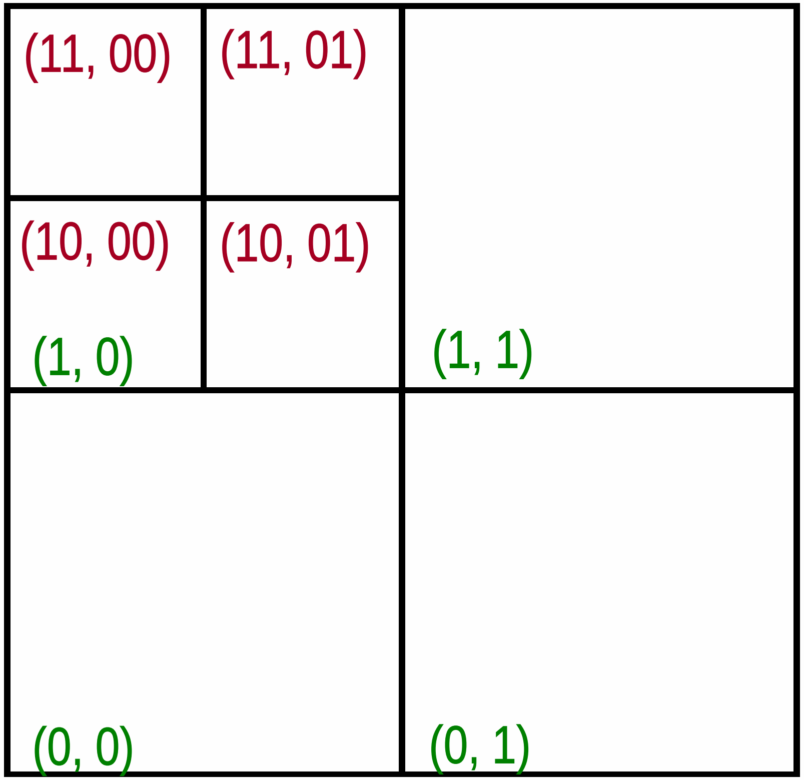
Only *k* out of *M* variables needed

$$\sum_{i=1}^{k} a_i(j) X_i + a_0(j) = 0$$

Hyperplane equation in parameter space

References: X Gan, A Liew, and H Yan, *BMC Bioinformatics*, 9:209, 2008;
X Gan, A Liew, and H Yan, "Representation and extraction of biclusters from data arrays," *US Patent* 7849088, 2010.

The Fast Hough Transform



(11, 00) (11, 01)

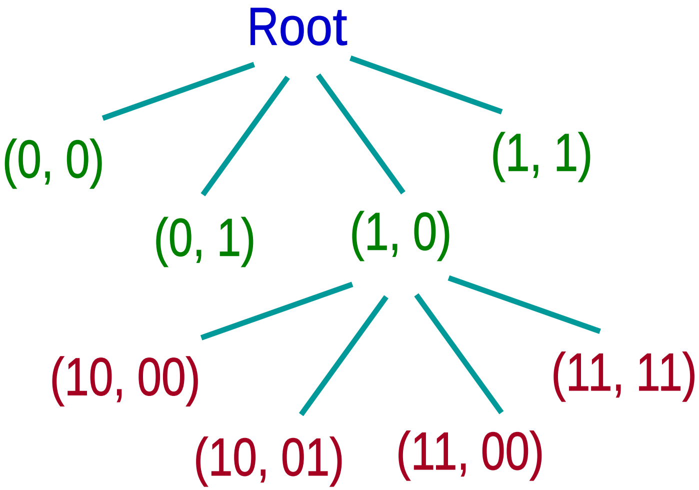
(10, 00) (10, 01)

(1, 0)

(1, 1)

(0, 0)

(0, 1)



Root

(0, 0) (0, 1) (1, 0) (1, 1)

(10, 00) (10, 01) (11, 00) (11, 11)

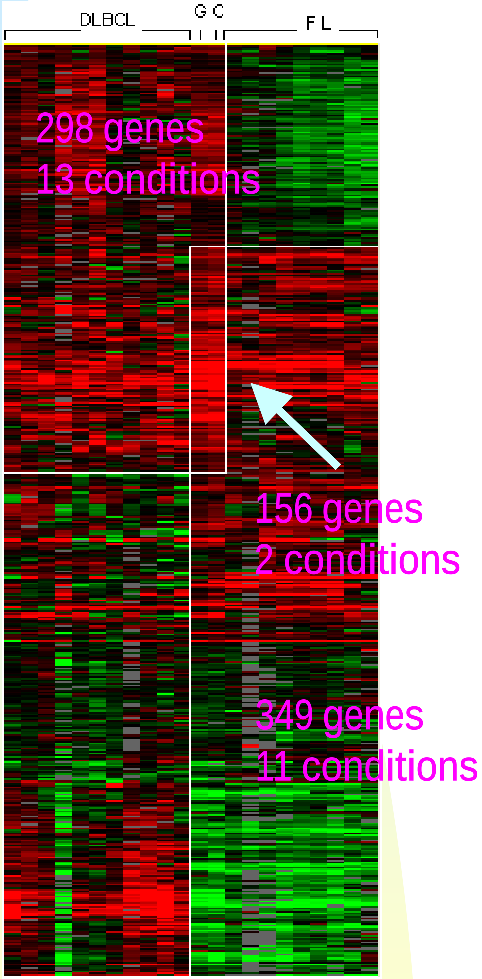
k-tree representation

Developed by:
Li, Levin and Le Master

$$a_0(j) + \sum_{i=1}^{k} a_i(j) C_i \leq r$$

Test for hyperplane and hypercube intersection

Cancer Diagnosis based on Co-clustering

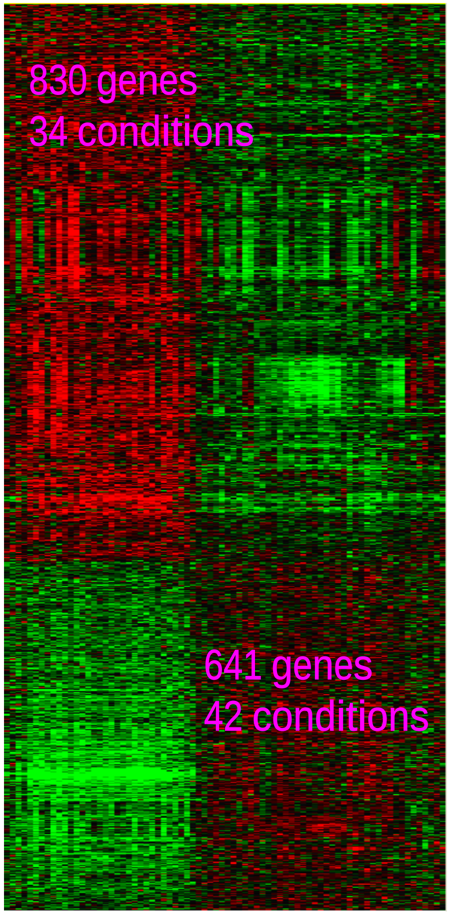


298 genes
13 conditions

156 genes
2 conditions

349 genes
11 conditions

Human Lymphoma Data

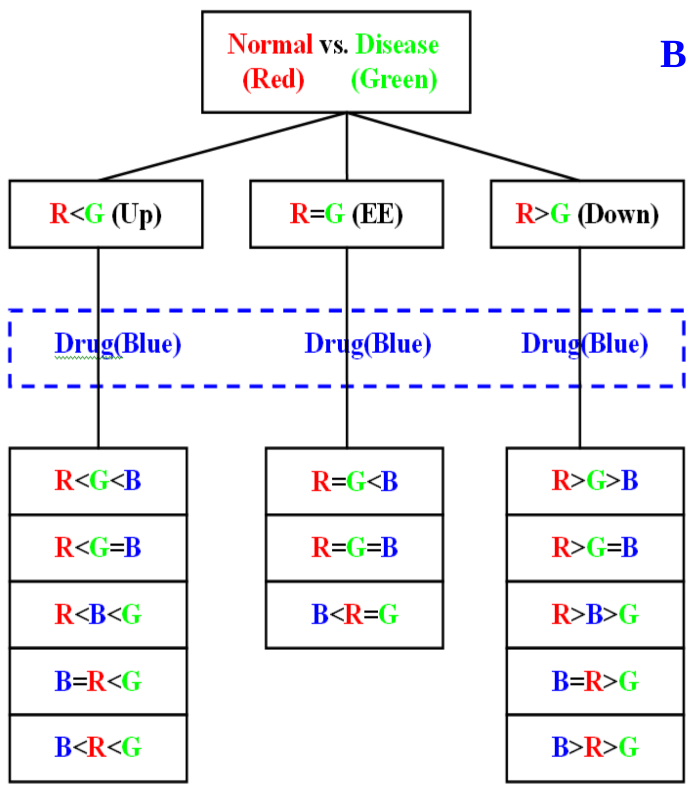

830 genes
34 conditions

641 genes
42 conditions

Breast Cancer Data

References:
X Gan, A Liew, and H Yan,
BMC Bioinformatics, 9:209, 2008;
X Gan, A Liew, and H Yan,
"Representation and extraction
of biclusters from data arrays,"
US Patent 7849088, 2010.

Drug Therapeutic Effect Assessment


Normal vs. Disease
(Red) (Green)

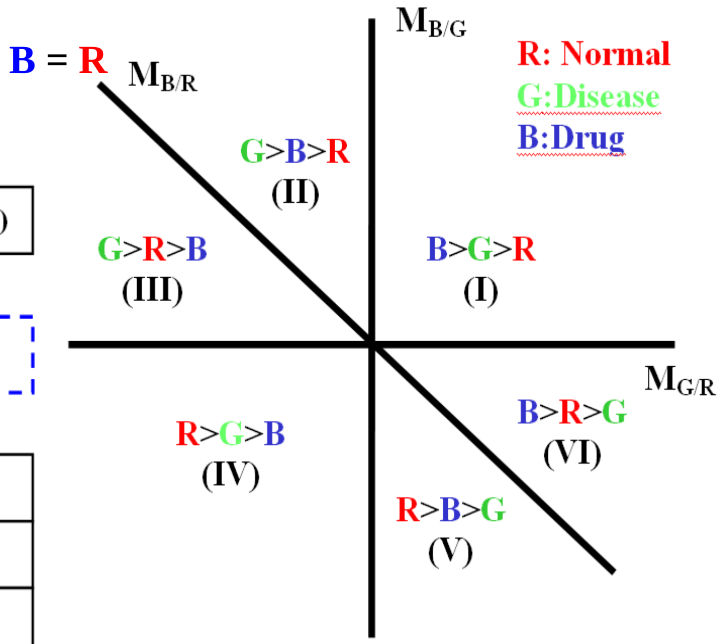
R<G (Up)
R=G (EE)
R>G (Down)

Drug(Blue)
Drug(Blue)
Drug(Blue)

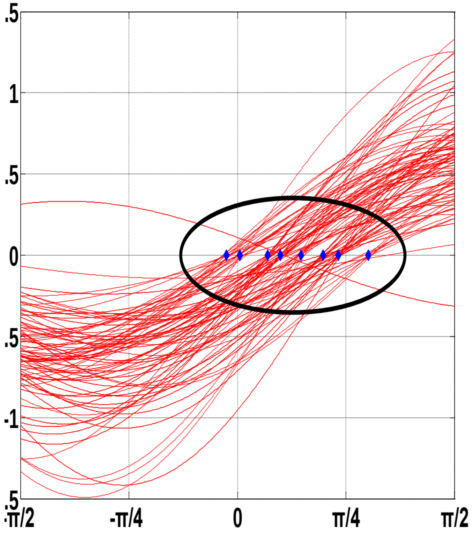
R<G<B
R<G=B
R<B<G
B=R<G
B<R<G

R=G<B
R=G=B
B<R=G

R>G>B
R>G=B
R>B>G
B=R>G
B>R>G



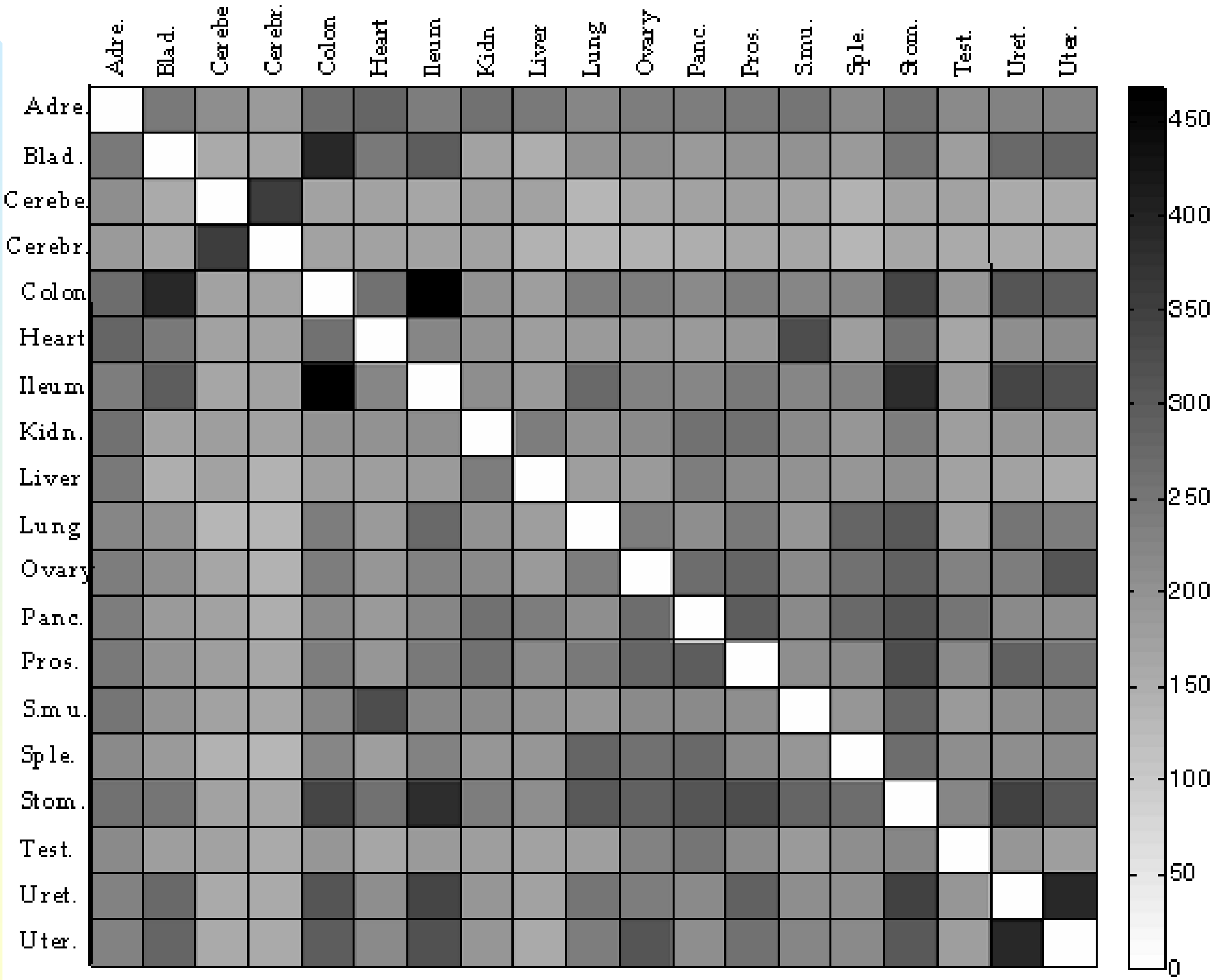
R: Normal
G:Disease
B:Drug



Slant axis: perfect drug
II, V: drug works, not enough
III, VI: drug works, too strong
I, IV: bad drug, side effect

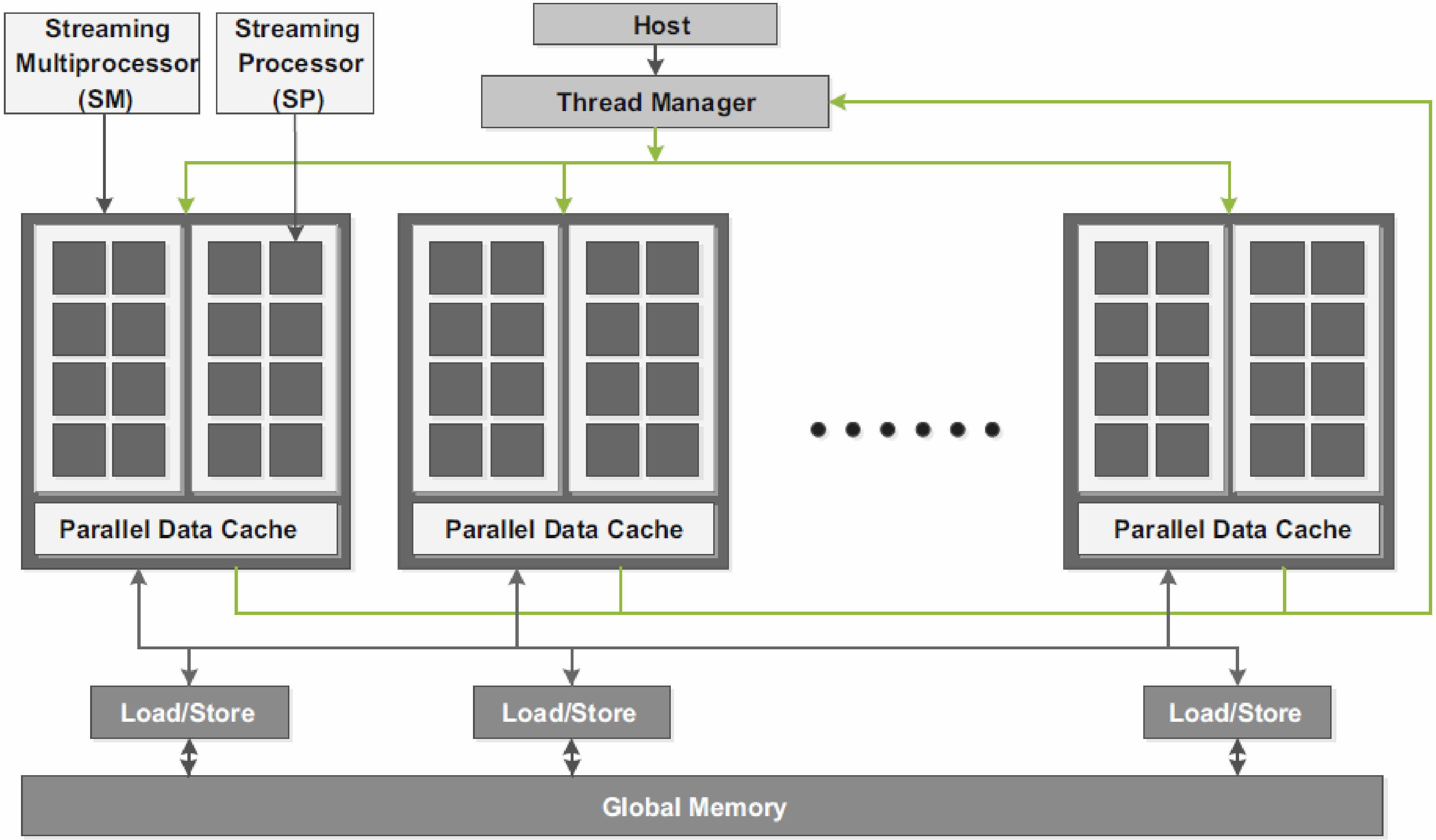
References: H. Zhao and H. Yan, *BMC Bioinformatics*, 8:256, 2007,
P. Tino, H. Zhao, and H. Yan, *IEEE Trans. Computational Biology & Bioinformatics*, 8:1093-1107, 2011.

Co-expressed Genes in Human Organs



Reference: H. Zhao, A. W. C. Liew, X. Xie, and H. Yan, *Journal of Theoretical Biology*, 251(3):264-274, 2008.

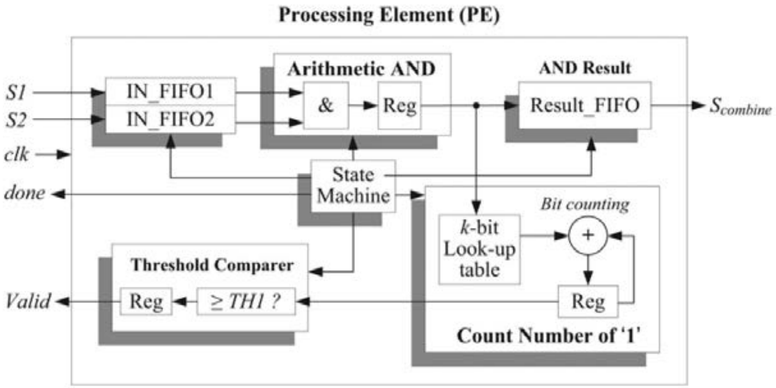
GPU Based Accelerators



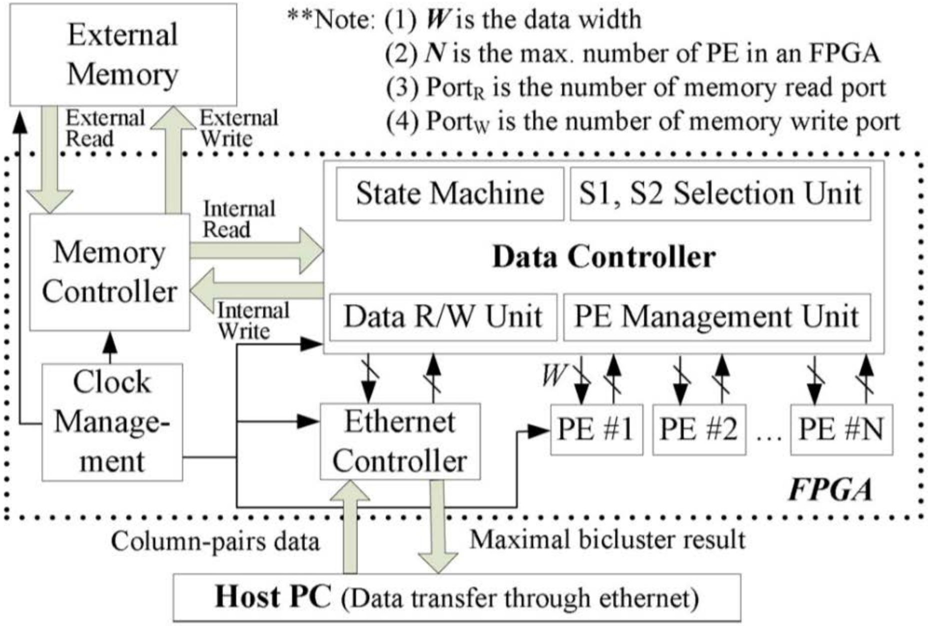
Reference: B Liu, Y Xin, RCC Cheung, and H Yan, *Neurocomputing*, 134:239–246, 2014.

FPGA Based Accelerators

Processing Element (PE)



FPGA Architecture



Reference:
B Liu, CW Yu, DZ Wang, RCC Cheung, H Yan,
IEEE T PDS, 25(10):2540-2550, 2014.

Complexity of HT for Many Variables

Assume:

Number of variables: *M*

Quantization level: *N*

→ Number of cells: M^N Too many for large *M*!

Solution:

Take 2 columns at a time

→ Analysis in column-pair spaces

Analysis in Column-pair Spaces

(1, 2) → sub-co-cluster
(1, 3) → sub-co-cluster
… …
(2, 3) → sub-co-cluster
(2, 4) → sub-co-cluster
… …
(3, 4) → sub-co-cluster
(3, 5) → sub-co-cluster
… ….

Merge sub-co-clusters
Form larger ones

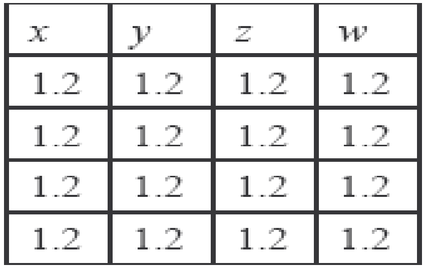
Limitations of Column-Pair Approach

Procedure:
- Consider each column pair
- Detect sub-co-clusters (with 2 columns and many rows)
- Merge sub-co-clusters

Limitations:
- Tow many column pairs for higher dimensional data
- Noise causes too many small sub-co-clusters
- Results depend on order of the merging process

Solution:
- Perform analysis in singular vector spaces

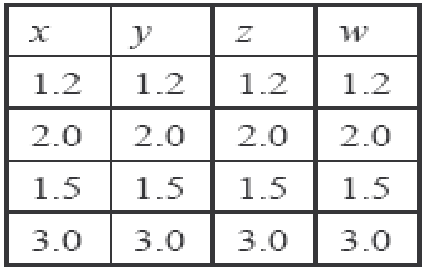
Co-clusters as Low-Rank Matrices



x	y	z	w
1.2	1.2	1.2	1.2
1.2	1.2	1.2	1.2
1.2	1.2	1.2	1.2
1.2	1.2	1.2	1.2

Constant

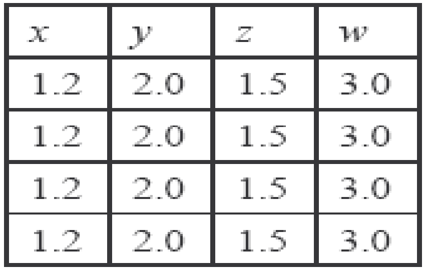
Rank = 1



x	y	z	w
1.2	1.2	1.2	1.2
2.0	2.0	2.0	2.0
1.5	1.5	1.5	1.5
3.0	3.0	3.0	3.0

Constant row

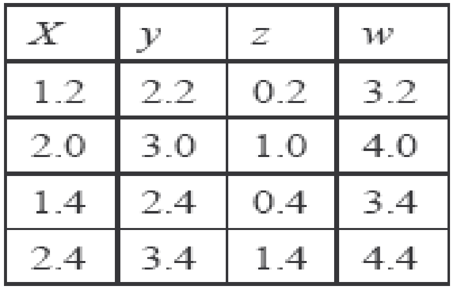
Rank = 1



x	y	z	w
1.2	2.0	1.5	3.0
1.2	2.0	1.5	3.0
1.2	2.0	1.5	3.0
1.2	2.0	1.5	3.0

Constant column

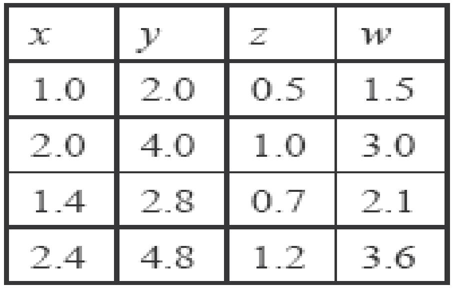
Rank = 1



X	y	z	w
1.2	2.2	0.2	3.2
2.0	3.0	1.0	4.0
1.4	2.4	0.4	3.4
2.4	3.4	1.4	4.4

Additive

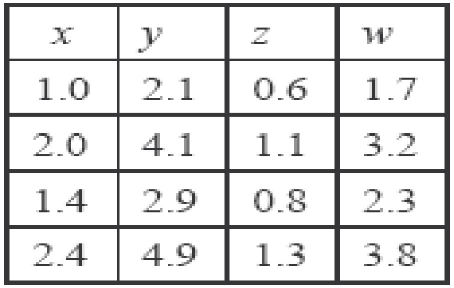
Rank = 2



x	y	z	w
1.0	2.0	0.5	1.5
2.0	4.0	1.0	3.0
1.4	2.8	0.7	2.1
2.4	4.8	1.2	3.6

Multiplicative

Rank = 1



x	y	z	w
1.0	2.1	0.6	1.7
2.0	4.1	1.1	3.2
1.4	2.9	0.8	2.3
2.4	4.9	1.3	3.8

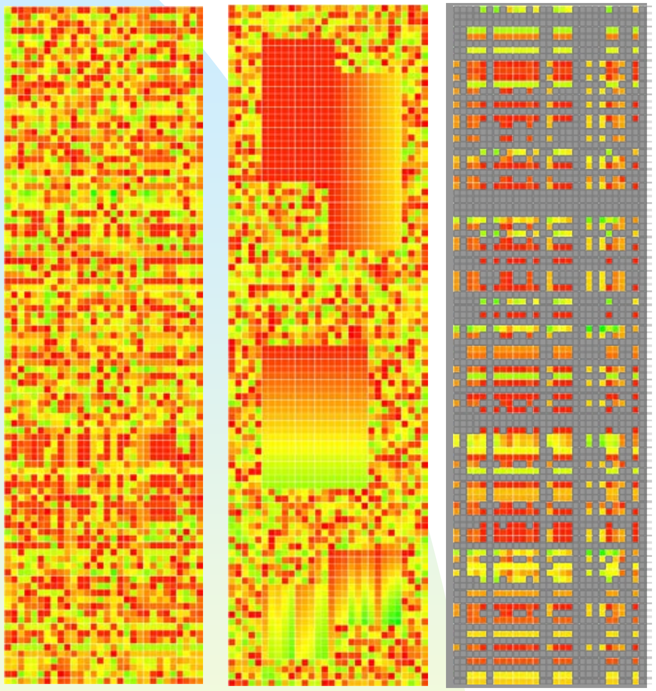
Linear

Rank = 2

Coherent pattern → Low rank matrix

Rank at most 2

Detection of Low-Rank Sub-matrices



XXXXXXXXXX XXXXXXXXXX XXXXXXXXXX

XAAXXXAXXX XAXAXXAXXX XAXXXAXXXA

XXXXXXXXXX XAXAXXAXXX XXXXXXXXXX

XXXXXXXXXX XXXXXXXXXX XAXXXAXXXA

XAAXXXAXXX XAXAXXAXXX XAXXXAXXXA

XXXXXXXXXX XXXXXXXXXX XXXXXXXXXX

XXXXXXXXXX XBXBXXBXXX XXBXBBXXXX

XAAXXXAXXX XXXXXXXXXX XXBXBBXXXX

XXXXXXXXXX XBXBXXBXXX XXXXXXXXXX

XXXXXXXXXX XBXBXXBXXX XXBXBBXXXX

Find locations of relevant elements

Together they form a low rank matrix

Decomposition of a 2D Co-cluster

$$\begin{aligned}
\mathbf{A}_c &= \mathbf{U}_c \Sigma_c \mathbf{V}_c^T \\
&= \begin{bmatrix} \mathbf{u}_1^c & \mathbf{u}_2^c \end{bmatrix} \begin{bmatrix} \sigma_1^c & 0 \\ 0 & \sigma_2^c \end{bmatrix} \begin{bmatrix} (\mathbf{v}_1^c)^T \\ (\mathbf{v}_2^c)^T \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{u}_1^c & \mathbf{u}_2^c \end{bmatrix} \begin{bmatrix} \sigma_1^c & 0 \\ 0 & \sigma_2^c \end{bmatrix} \begin{bmatrix} v_{11}^c & v_{12}^c & \cdots & v_{1d}^c \\ v_{21}^c & v_{22}^c & \cdots & v_{2d}^c \end{bmatrix} \\
&= \begin{bmatrix} \sigma_1^c v_{11}^c \mathbf{u}_1^c + \sigma_2^c v_{21}^c \mathbf{u}_2^c & \sigma_1^c v_{12}^c \mathbf{u}_1^c + \sigma_2^c v_{22}^c \mathbf{u}_2^c & \cdots & \sigma_1^c v_{1d}^c \mathbf{u}_1^c + \sigma_2^c v_{2d}^c \mathbf{u}_2^c \end{bmatrix}
\end{aligned}$$

Reference:
H. Zhao, D. D. Wang, L. Chen, X. Liu, and H. Yan, *PLoS ONE*, 11(9): e0162293:1-27, 2016.

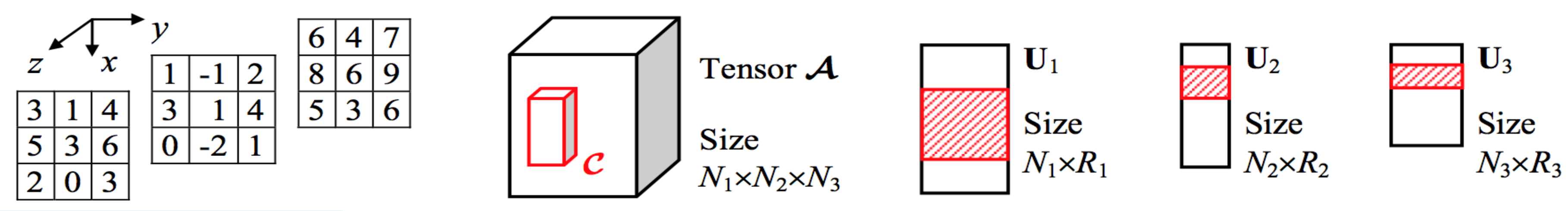
$$\mathbf{a}_j^c = \alpha_{j1} \mathbf{u}_1^c + \alpha_{j2} \mathbf{u}_2^c$$

$$\mathbf{a}_i^c = \beta_{i1} \mathbf{v}_1^c + \beta_{i2} \mathbf{v}_2^c$$

$$\begin{cases} \alpha_{j1} u_{11}^c + \alpha_{j2} u_{12}^c = a_{1j}^c \\ \alpha_{j1} u_{21}^c + \alpha_{j2} u_{22}^c = a_{2j}^c \\ \qquad \vdots \\ \alpha_{j1} u_{s1}^c + \alpha_{j2} u_{s2}^c = a_{sj}^c \end{cases}$$

$$\begin{cases} \beta_{i1} v_{11}^c + \beta_{i2} v_{12}^c = a_{i1}^c \\ \beta_{i1} v_{21}^c + \beta_{i2} v_{22}^c = a_{i2}^c \\ \qquad \vdots \\ \beta_{i1} v_{t1}^c + \beta_{i2} v_{t2}^c = a_{it}^c \end{cases}$$

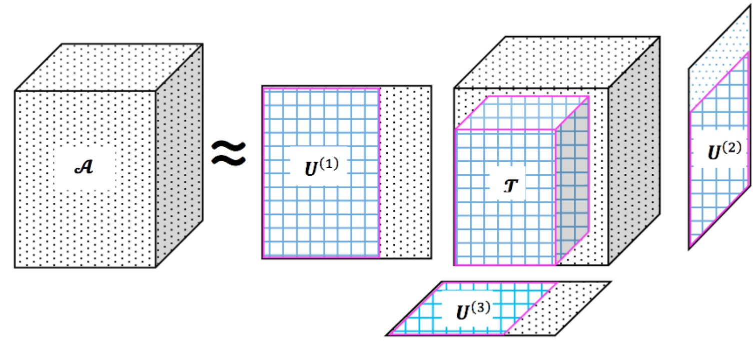
Decomposition of a 3D Co-cluster



3	1	4
5	3	6
2	0	3

1	-1	2
3	1	4
0	-2	1

6	4	7
8	6	9
5	3	6



Truncated HOSVD for tensor approximation

Rank $(r_1, r_2, \dots, r_N)$

$$\mathscr{A} = \mathscr{T} \times_1 \mathbf{U}^{(1)} \dots \times_N \mathbf{U}^{(N)}$$

$$\mathbf{A}_{(n)} = \mathbf{U}^{(n)} \Sigma^{(n)} (\mathbf{V}^{(n)})^T, \quad n = 1, \dots, N$$

Hyperplane Structure

- If $\boldsymbol{C}$ is a coherent matrix, then $rank(\boldsymbol{C}) \leq 2$.
- If a coherent pattern $\boldsymbol{A}_c = \boldsymbol{U}_c \boldsymbol{\Lambda}_c \boldsymbol{V}_c^T$, then the column vectors of $\boldsymbol{U}_c$ and $\boldsymbol{V}_c$ are linearly dependent, i.e.

$$\begin{cases} k_{1u}\boldsymbol{u}_{1c} + k_{2u}\boldsymbol{u}_{2c} + b_u\boldsymbol{1} = 0 \\ k_{1v}\boldsymbol{v}_{1c} + k_{2v}\boldsymbol{v}_{2c} + b_v\boldsymbol{1} = 0 \end{cases}$$

- If a coherent pattern $\mathcal{A}_c = \mathcal{T}_c \times_1 \boldsymbol{U}_{1c} \times_2 \boldsymbol{U}_{2c} \cdots \times_N \boldsymbol{U}_{Nc}$, then the column vectors of $\boldsymbol{U}_{ic}$ are linearly dependent, i.e.

$$\begin{cases} k_{11}\boldsymbol{u}_{1c}^{1} + k_{12}\boldsymbol{u}_{1c}^{2} + \cdots k_{1r_1}\boldsymbol{u}_{1c}^{r_1} + b_1\boldsymbol{1} = 0 \\ \vdots \\ k_{N1}\boldsymbol{u}_{Nc}^{1} + k_{N2}\boldsymbol{u}_{Nc}^{2} + \cdots k_{Nr_N}\boldsymbol{u}_{1c}^{r_N} + b_N\boldsymbol{1} = 0 \end{cases}$$

Co-cluster Scoring Function

2D Data

$$S(\mathbf{I},\mathbf{J}) = \min_{i\in\mathbf{I},\, j\in\mathbf{J}} \left(S_{\mathbf{I}j}, S_{i\mathbf{J}}\right) = \min_{i\in\mathbf{I},\, j\in\mathbf{J}} \left[1 - \frac{1}{|\mathbf{J}|-1} \sum_{q\neq j, q\in\mathbf{J}} \rho\left(\mathbf{a}_{\mathbf{I}j}, \mathbf{a}_{\mathbf{I}q}\right), 1 - \frac{1}{|\mathbf{I}|-1} \sum_{q\neq i, q\in\mathbf{I}} \rho\left(\mathbf{a}_{i\mathbf{J}}, \mathbf{a}_{q\mathbf{J}}\right)\right]$$

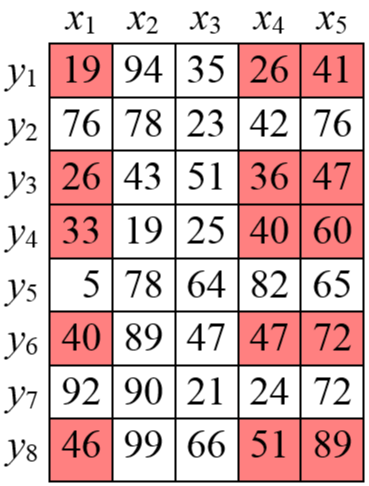
N-D Data

$$S(\mathbf{I}^{(1)}, \mathbf{I}^{(2)}, \cdots, \mathbf{I}^{(N)}) = \min_{i_1\in\mathbf{I}^{(1)}, i_2\in\mathbf{I}^{(2)}, \cdots, i_N\in\mathbf{I}^{(N)}} \left(S_{i_1\mathbf{I}^{(2)}\ldots\mathbf{I}^{(N)}}, S_{\mathbf{I}^{(1)}i_2\mathbf{I}^{(3)}\ldots\mathbf{I}^{(N)}}, \cdots S_{\mathbf{I}^{(1)}\mathbf{I}^{(2)}\ldots i_N}\right)$$

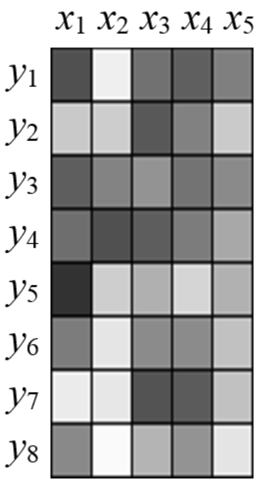
Pair row and column indices

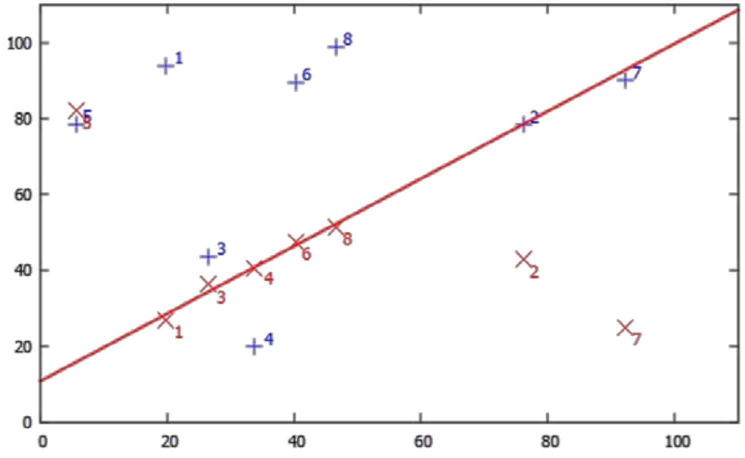
Check coherence and filter out noisy patterns

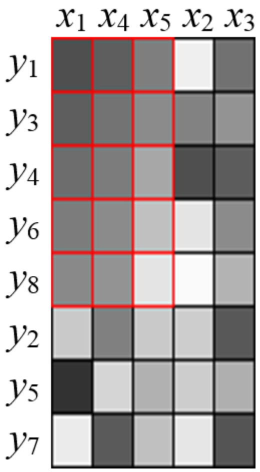
2D Co-clusters in Singular Vector Spaces

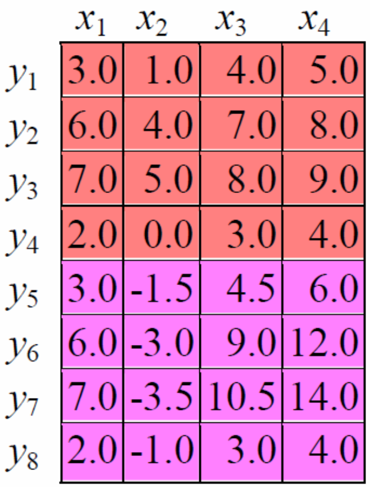


	x_1	x_2	x_3	x_4	x_5
y_1	19	94	35	26	41
y_2	76	78	23	42	76
y_3	26	43	51	36	47
y_4	33	19	25	40	60
y_5	5	78	64	82	65
y_6	40	89	47	47	72
y_7	92	90	21	24	72
y_8	46	99	66	51	89





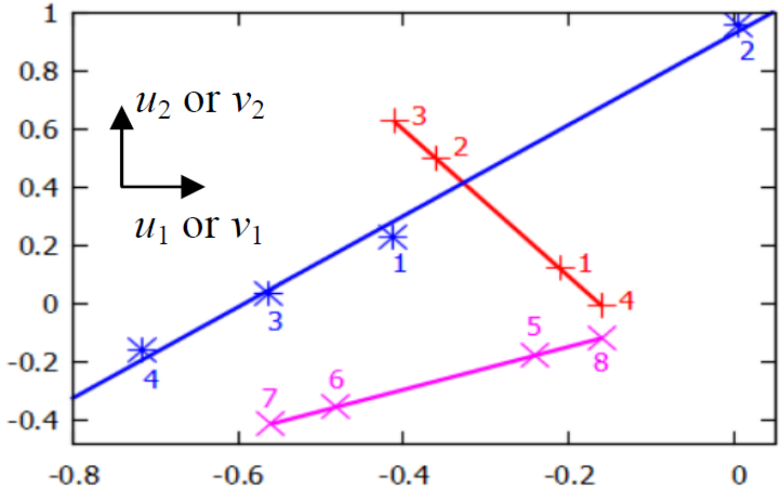




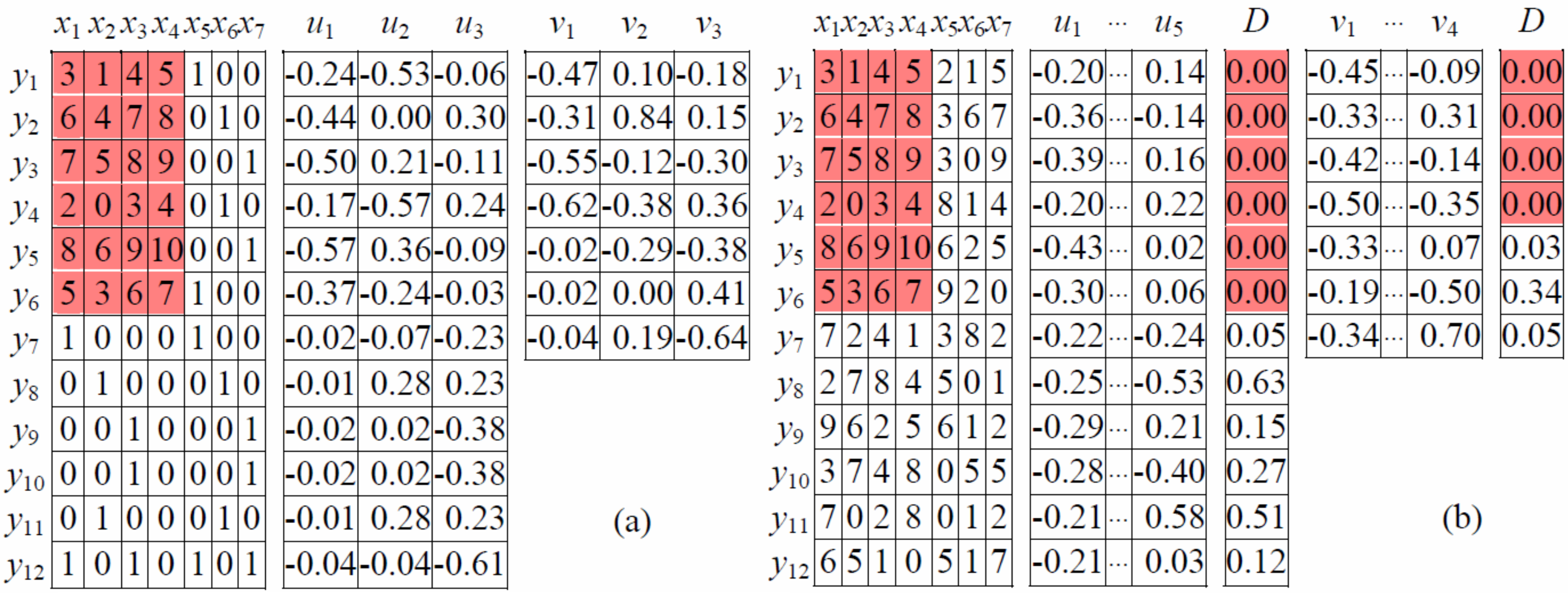
	x_1	x_2	x_3	x_4
y_1	3.0	1.0	4.0	5.0
y_2	6.0	4.0	7.0	8.0
y_3	7.0	5.0	8.0	9.0
y_4	2.0	0.0	3.0	4.0
y_5	3.0	-1.5	4.5	6.0
y_6	6.0	-3.0	9.0	12.0
y_7	7.0	-3.5	10.5	14.0
y_8	2.0	-1.0	3.0	4.0

u_1	u_2
-0.21	0.12
-0.36	0.50
-0.41	0.63
-0.16	-0.01
-0.24	-0.18
-0.48	-0.35
-0.56	-0.41
-0.16	-0.12

v_1	v_2
-0.41	0.23
0.01	0.96
-0.56	0.04
-0.72	-0.16



3D Co-clusters in Singular Vector Spaces

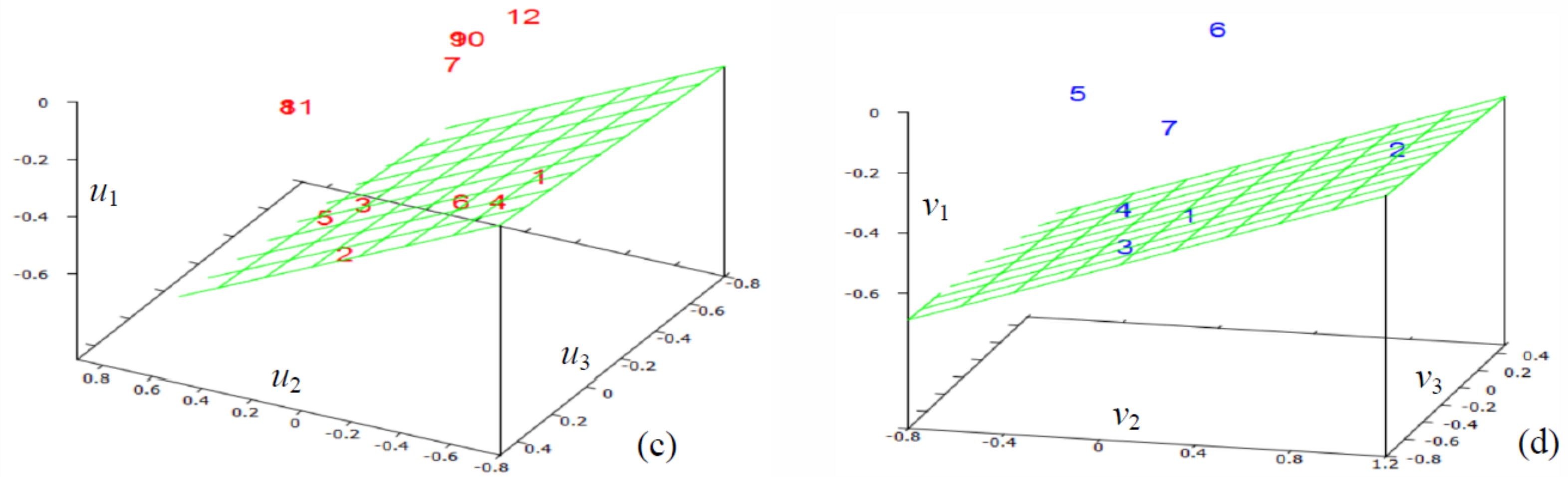


	x_1	x_2	x_3	x_4	x_5	x_6	x_7	u_1	u_2	u_3	v_1	v_2	v_3
y_1	3	1	4	5	1	0	0	-0.24	-0.53	-0.06	-0.47	0.10	-0.18
y_2	6	4	7	8	0	1	0	-0.44	0.00	0.30	-0.31	0.84	0.15
y_3	7	5	8	9	0	0	1	-0.50	0.21	-0.11	-0.55	-0.12	-0.30
y_4	2	0	3	4	0	1	0	-0.17	-0.57	0.24	-0.62	-0.38	0.36
y_5	8	6	9	10	0	0	1	-0.57	0.36	-0.09	-0.02	-0.29	-0.38
y_6	5	3	6	7	1	0	0	-0.37	-0.24	-0.03	-0.02	0.00	0.41
y_7	1	0	0	0	1	0	0	-0.02	-0.07	-0.23	-0.04	0.19	-0.64
y_8	0	1	0	0	0	1	0	-0.01	0.28	0.23			
y_9	0	0	1	0	0	0	1	-0.02	0.02	-0.38			
y_{10}	0	0	1	0	0	0	1	-0.02	0.02	-0.38			
y_{11}	0	1	0	0	0	1	0	-0.01	0.28	0.23			
y_{12}	1	0	1	0	1	0	1	-0.04	-0.04	-0.61			

(a)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	u_1	···	u_5	D	v_1	···	v_4	D
y_1	3	1	4	5	2	1	5	-0.20	···	0.14	0.00	-0.45	···	-0.09	0.00
y_2	6	4	7	8	3	6	7	-0.36	···	-0.14	0.00	-0.33	···	0.31	0.00
y_3	7	5	8	9	3	0	9	-0.39	···	0.16	0.00	-0.42	···	-0.14	0.00
y_4	2	0	3	4	8	1	4	-0.20	···	0.22	0.00	-0.50	···	-0.35	0.00
y_5	8	6	9	10	6	2	5	-0.43	···	0.02	0.00	-0.33	···	0.07	0.03
y_6	5	3	6	7	9	2	0	-0.30	···	0.06	0.00	-0.19	···	-0.50	0.34
y_7	7	2	4	1	3	8	2	-0.22	···	-0.24	0.05	-0.34	···	0.70	0.05
y_8	2	7	8	4	5	0	1	-0.25	···	-0.53	0.63				
y_9	9	6	2	5	6	1	2	-0.29	···	0.21	0.15				
y_{10}	3	7	4	8	0	5	5	-0.28	···	-0.40	0.27				
y_{11}	7	0	2	8	0	1	2	-0.21	···	0.58	0.51				
y_{12}	6	5	1	0	5	1	7	-0.21	···	0.03	0.12				

(b)



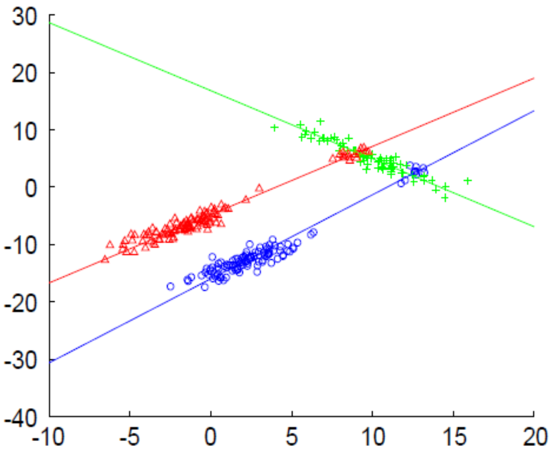
(c)

(d)

Detection of Hyperplanes

Detect linear patterns in data sets

- Scaling of the variable
- Generation of the starting hyperplanes
- Initialization of the groups
- Iterative refinement
- Resampling



Scale the variables for $\boldsymbol{i = 1, \cdots, d}$,

$$\widetilde{\boldsymbol{y}}_i = \frac{y_i}{s_i} \quad (\boldsymbol{i = 1, \cdots, d}), \boldsymbol{s}_i = \sqrt{\frac{1}{n-1}(\boldsymbol{y}_\mathrm{i} - \overline{\boldsymbol{y}}_\mathrm{i})^T(\boldsymbol{y}_\mathrm{i} - \overline{\boldsymbol{y}}_\mathrm{i})}$$

Randomly select $\boldsymbol{K}$ random sub-samples of size $\boldsymbol{d}$, $\boldsymbol{G^0} = \{\boldsymbol{g_1^0, \cdots, g_K^0}\}$

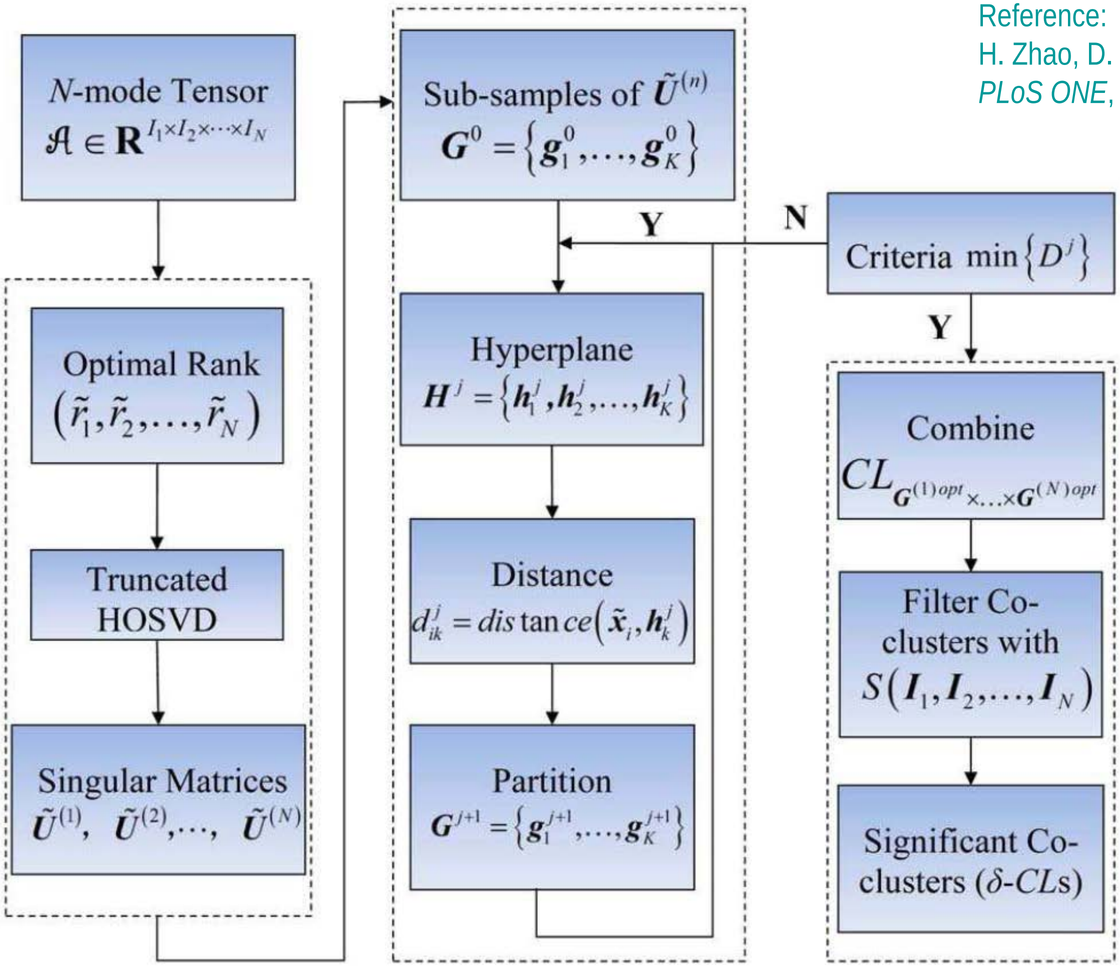
```
Iterative procedure

    Initializing $K$ hyperplanes
        $h_k^j(\widehat{\alpha}_k, \widehat{\beta}_k) = \{\widetilde{x} | \alpha_k^T \widetilde{x} = \beta_k, \widetilde{x} \in g_k^j, \|\alpha_k\| = 1\} \ (k = 1, \cdots, K)$
    Compute $d_{ik}^j = distance(\widetilde{x}_i, h_k^j) = |\widehat{\alpha}_k^T \widetilde{x}_i - \widehat{\beta}_k|$
    Forming $K$ groups for $n$ samples such that
        $\widetilde{x}_i \in g_k^{j+1}$ if $k = argmin_k(d_{ik}^j)$
    Computing the cost function $D^j = \sum_{k=1}^{K} \sum_{\widetilde{x}_i \in g_k^{j+1}} d_{ik}^j$

End
```

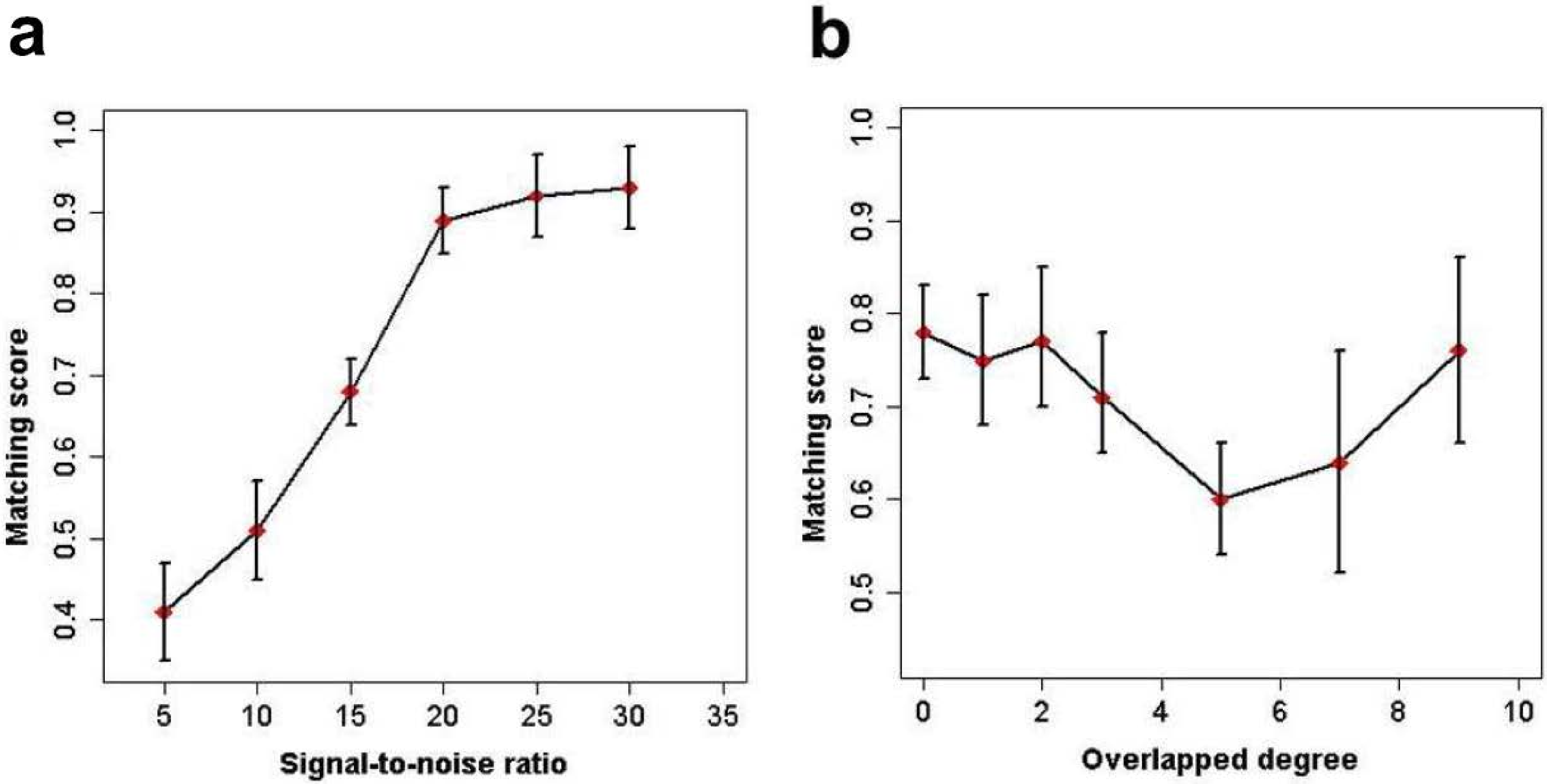
Reference:
H. Zhao, D. D. Wang, L. Chen, X. Liu, and H. Yan, *PLoS ONE*, 11(9): e0162293:1-27, 2016.

Coherent Pattern Detection Algorithm

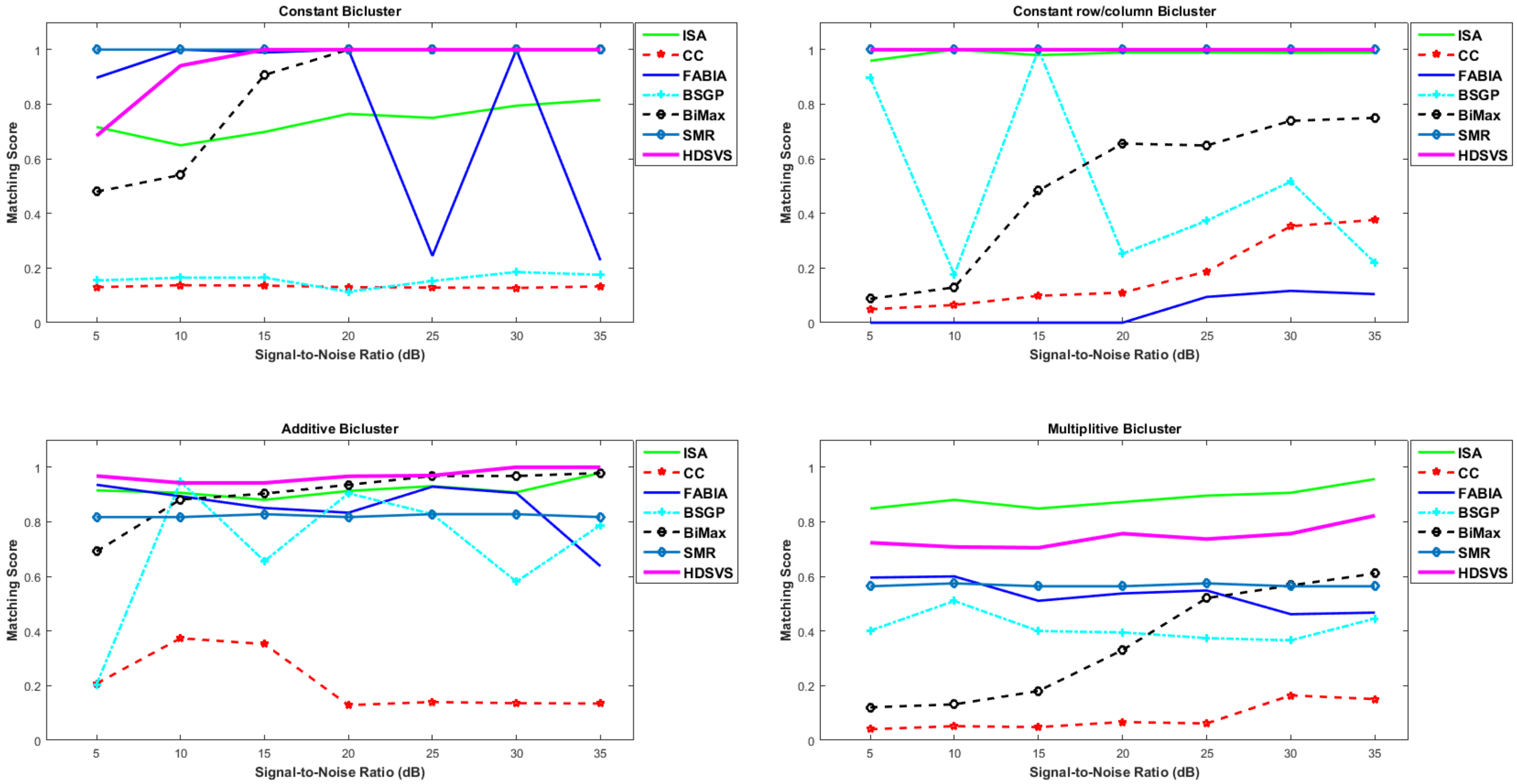


Reference:
H. Zhao, D. D. Wang, L. Chen, X. Liu, and H. Yan,
PLoS ONE, 11(9): e0162293:1-27, 2016.

Experiments on Simulated Triclusters



Comparison with Other Methods

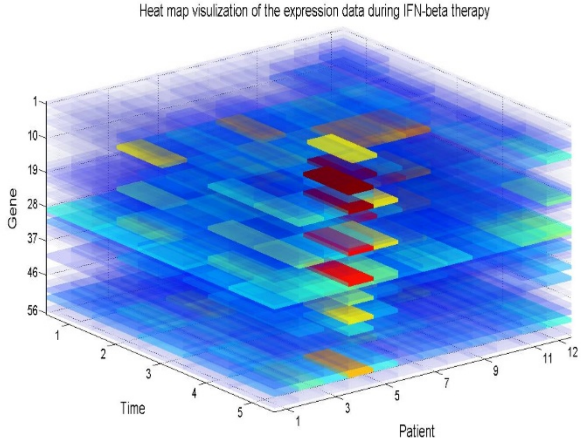


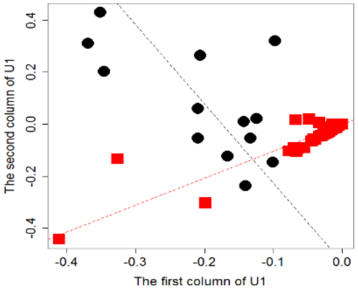
Matrix size: 500 x 200, Bicluster size: 50 x 50

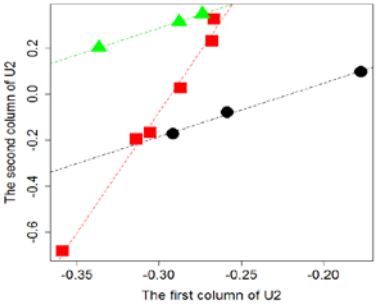
Example of 3D Data

A higher-order time series dataset about genomic expression of multiple sclerosis patients after IFN-b injection treatment.

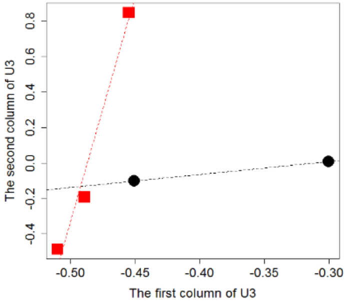
- Patient: Twelve patients
- Time: EDTA blood samples from patients before baseline as well as 2 days, 1 month, 1 year, and 2 years after the initiation of IFN-beta therapy.
- Gene: 56 significant genes involved in IFN-related pathways



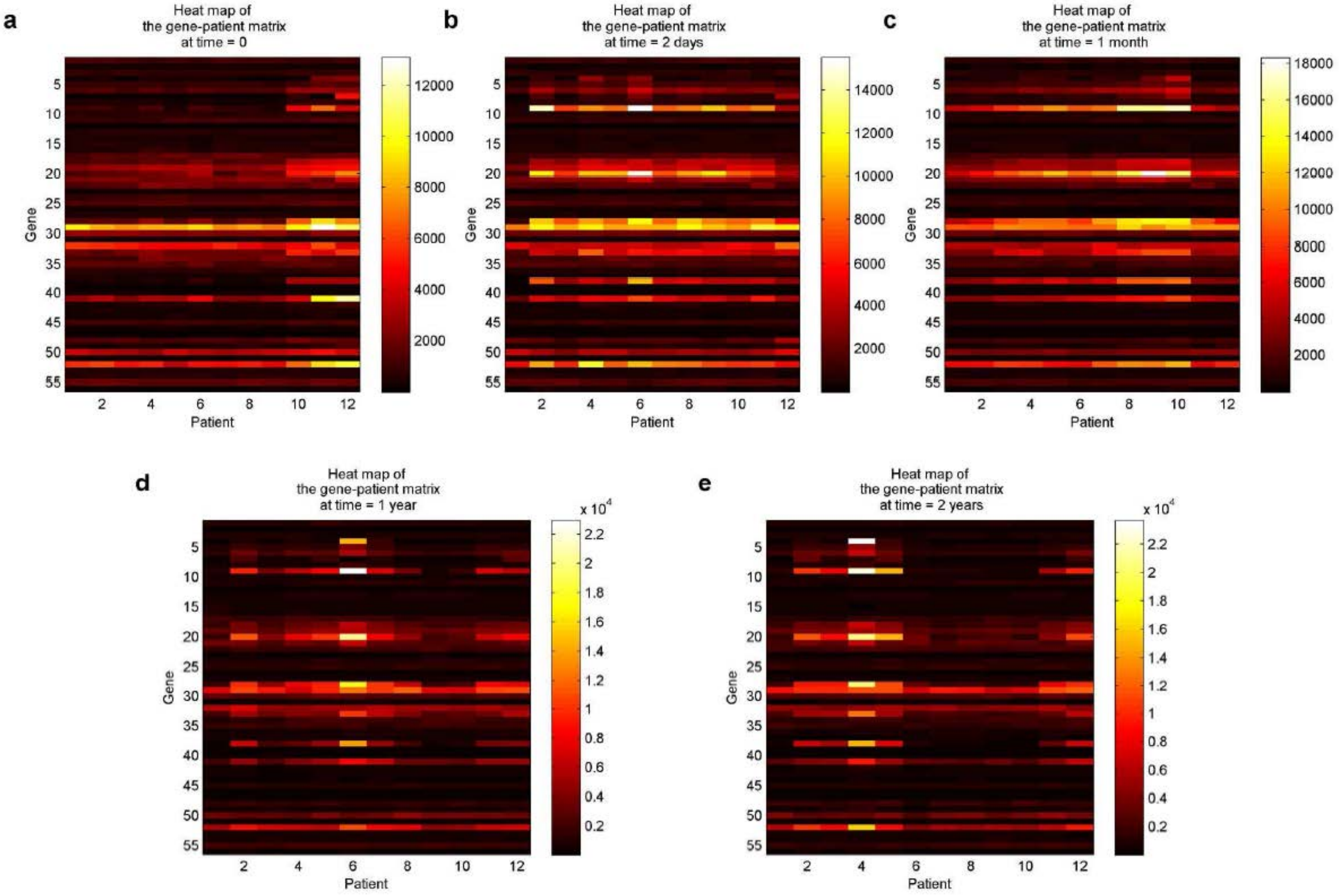




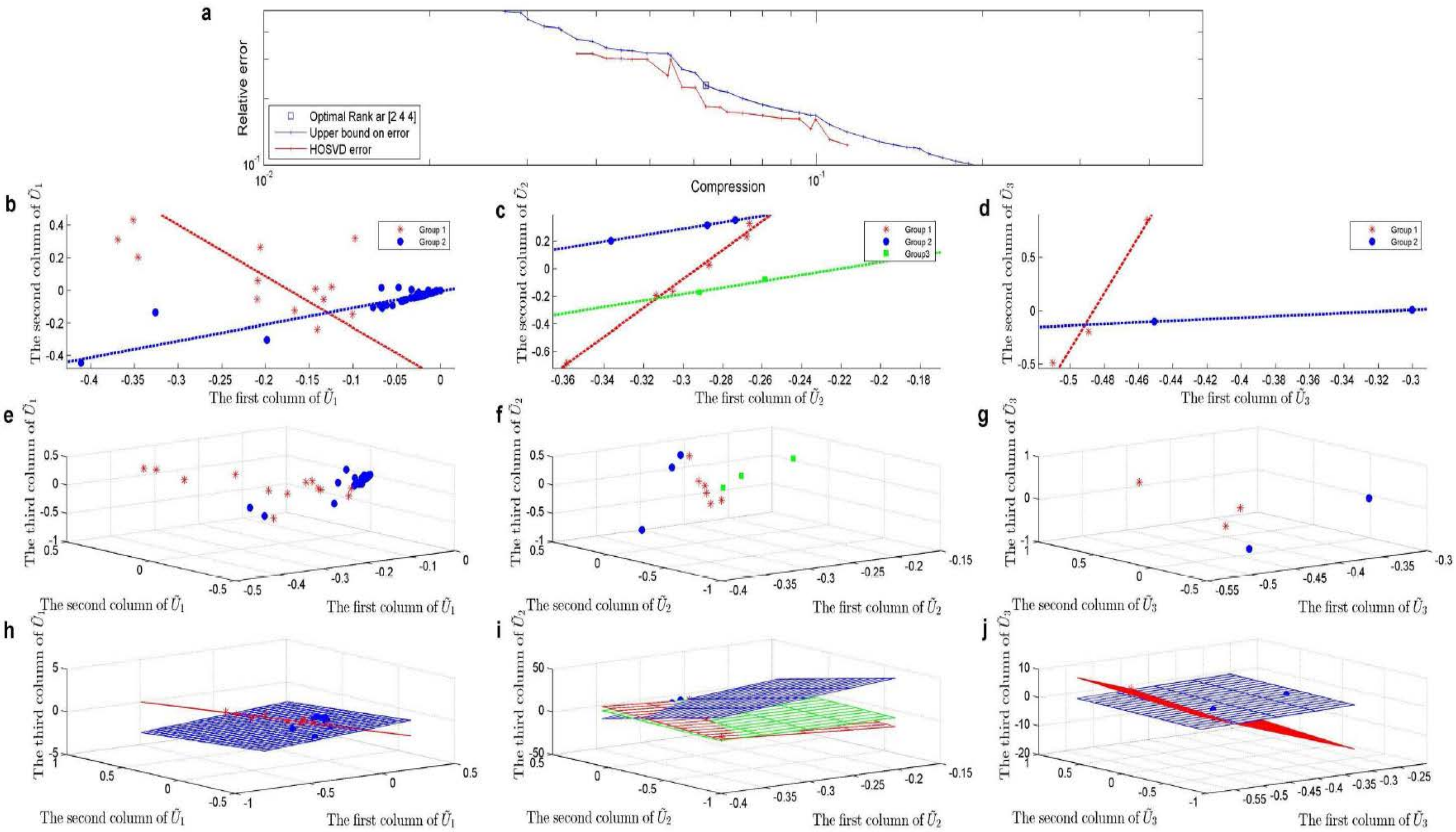
Linear structures along 3 directions



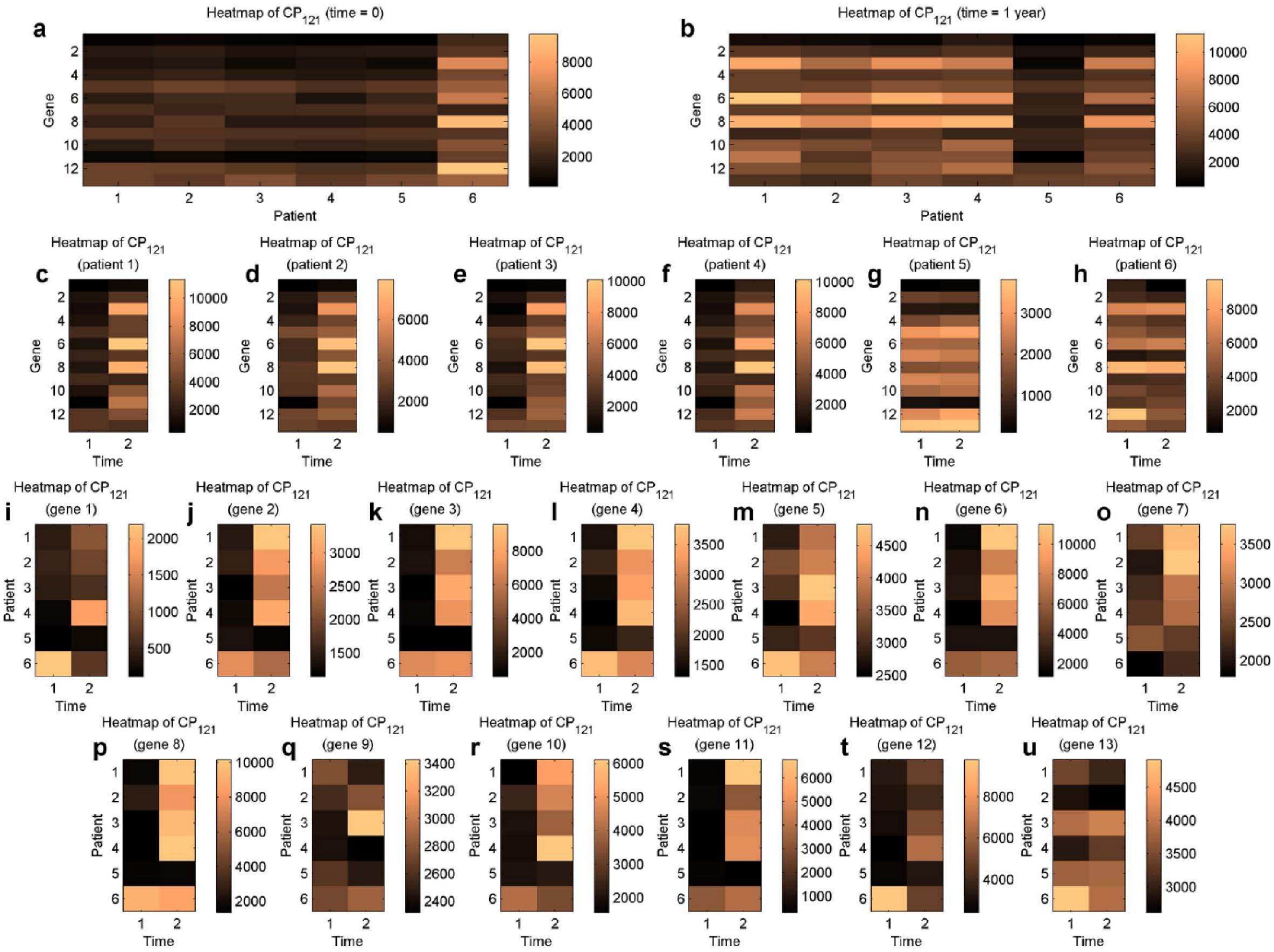
2D Slices of Sclerosis Data



Analysis of Sclerosis Data



Example of Co-cluster in Sclerosis Data



GO & Pathway Analysis of Sclerosis Data

13 genes: CXCL10, EIF2AK2, IFIT1, IRF7, IRF9, ISG15, ISG20, MX1, NFKB1, OAS1, RSAD2, STAT1, TLR8;

6 patients: with the common clinical features such as the shorter disease duration, the lower EDSS scores, and relapses prior to 1 year;

2 time points: baseline and 1 year

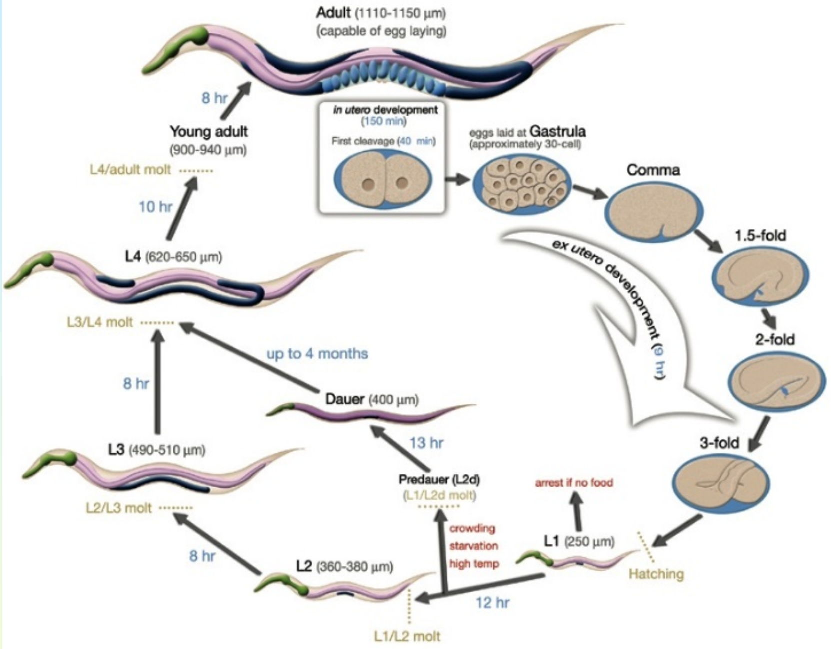
Table 1 Biological process of 13 annotated genes in CP_{121}.

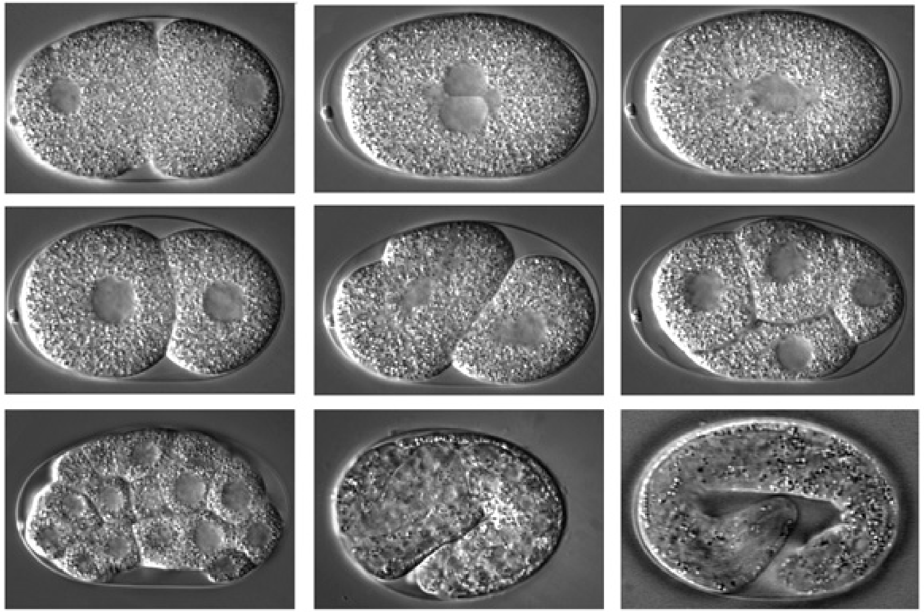
GO term	Description	P-value	Enrichment
GO: 0002252	immune effector process	8.81E-5	2.07 (56,25,13,12)
GO: 0051607	defense response to virus	1.17E-4	2.26 (56,21,13,11)
GO: 0045069	regulation of viral genome replication	2.12E-4	3.35 (56,9,13,7)
GO: 0045071	negative regulation of viral genome replication	2.12E-4	3.35 (56,9,13,7)
GO: 0009615	response to virus	2.77E-4	1.91 (56,27,13,12)
GO: 0050792	regulation of viral process	6.28E-4	3.02 (56,10,13,7)
GO: 0048525	negative regulation of viral process	6.28E-4	3.02 (56,10,13,7)

Table 2 Genomic pathway of 13 annotated genes in CP_{121}

Pathway annotated	Term	P-value	Adjusted P-value
BIOCARTA	Bone remodelling	2.3E-4	1.0E-2
BIOCARTA	IFN alpha signaling pathway	1.9E-2	3.5E-1
BIOCARTA	Double stranded RNA induced gene expression	1.9E-2	3.5E-1
BIOCARTA	Inactivation of Gsk3 by AKT causes accumulation of b-catenin in alveolar macrophages	5.9E-2	6.0E-1
BIOCARTA	Toll-like Receptor Pathway	7.5E-2	5.9E-1
KEGG_PATHWAY	Toll-like receptor signaling pathway	2.1E-6	4.3E-5
KEGG_PATHWAY	RIG-I-like receptor signaling pathway	5.1E-5	5.1E-4
KEGG_PATHWAY	Cytosolic DNA-sensing pathway	1.7E-3	1.1E-2
KEGG_PATHWAY	Chemokine signaling pathway	1.8E-2	8.8E-2
KEGG_PATHWAY	Pancreatic cancer	8.2E-2	2.9E-1

C. Elegans Life Cycle



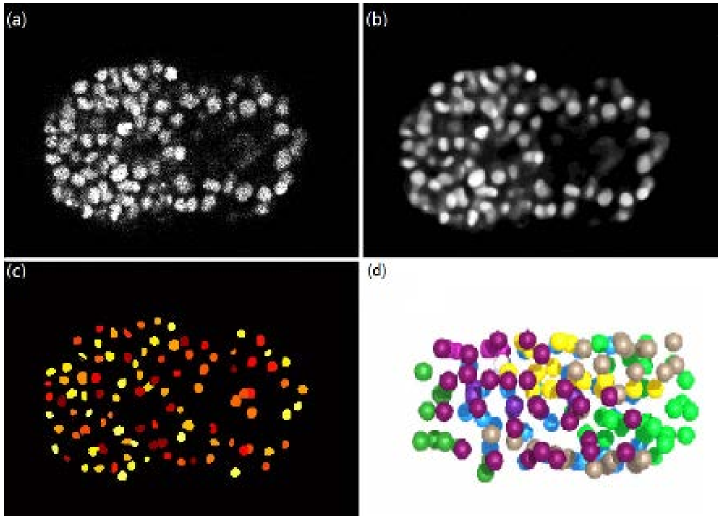


Picture from
www.wormatlas.org/ver1/handbook/anatomyintro
/anatomyintro.htm

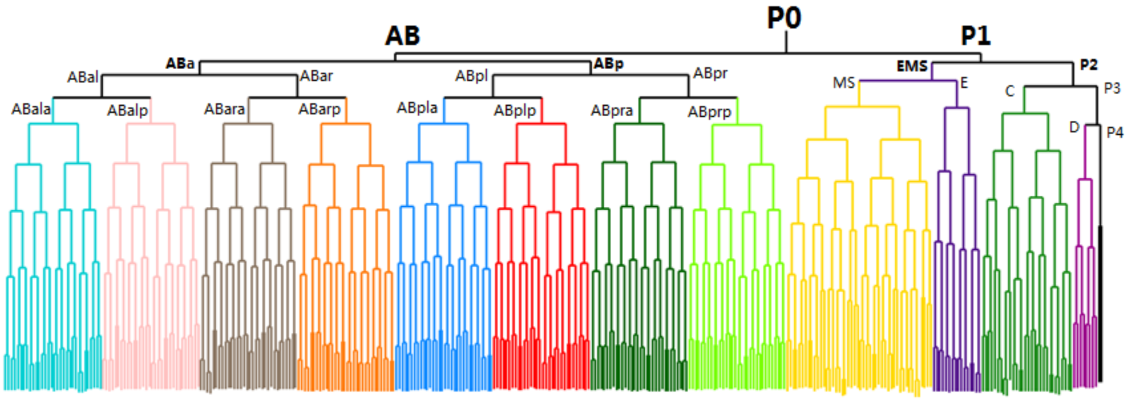
Picture from
http://www.mun.ca/biology/scarr/4241_Devo_
Caenorhabditis%20elegans%20devo.jpg

C. Elegans Cell Division

Image/video acquisition



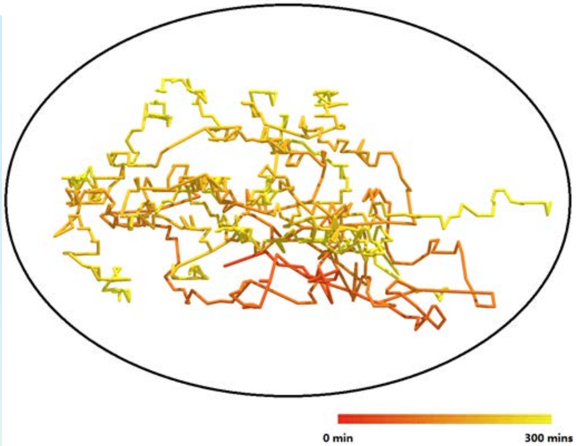
Analysis of image/video (Tera bytes of data)

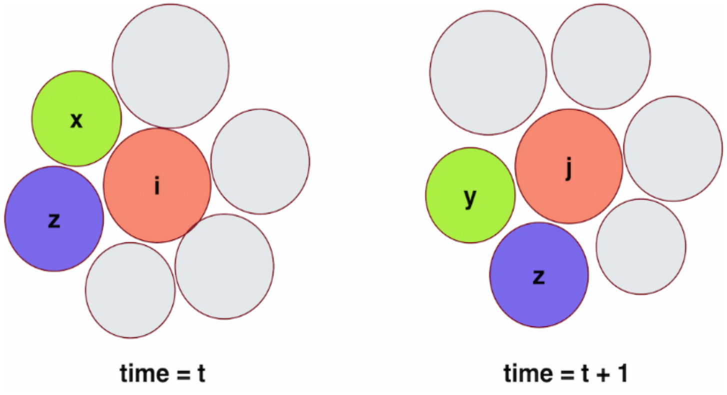


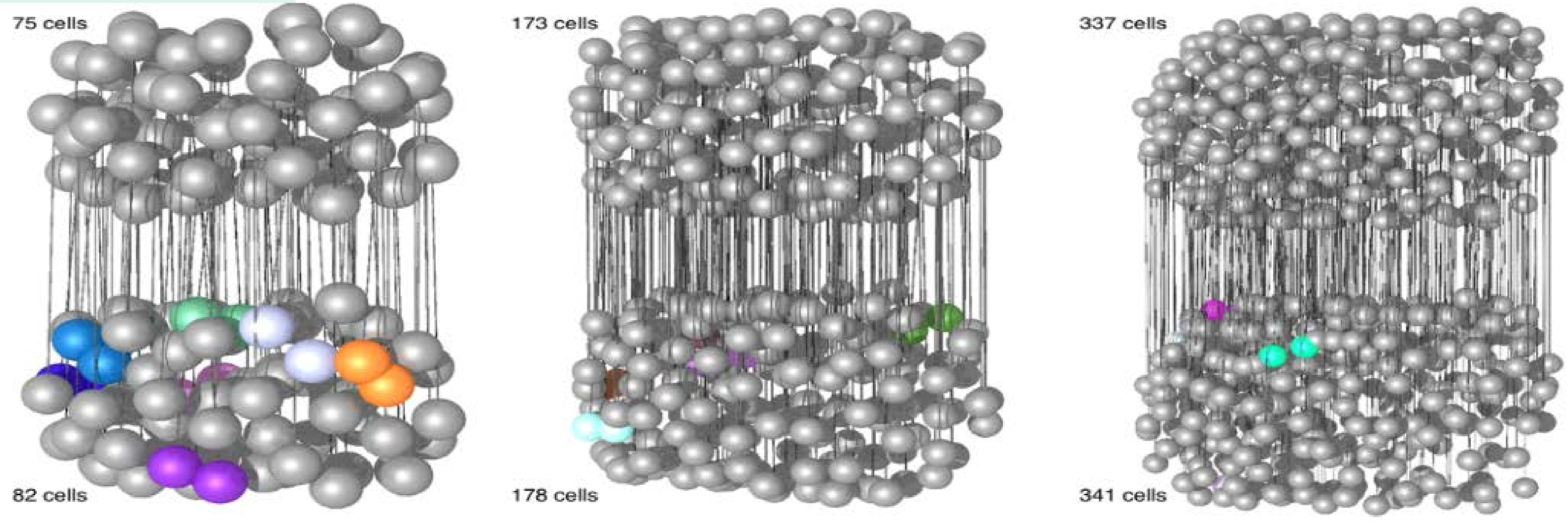
Images to lineage tree

References:
L Chen, LLH Chan, Z Zhao, H Yan, *BMC Bioinformatics*, 14:328, 2013.
J Cao, MK Wong, Z Zhao, and H Yan, *BMC Bioinformatics*, 20:176, 2019.

Cell Matching and Matching

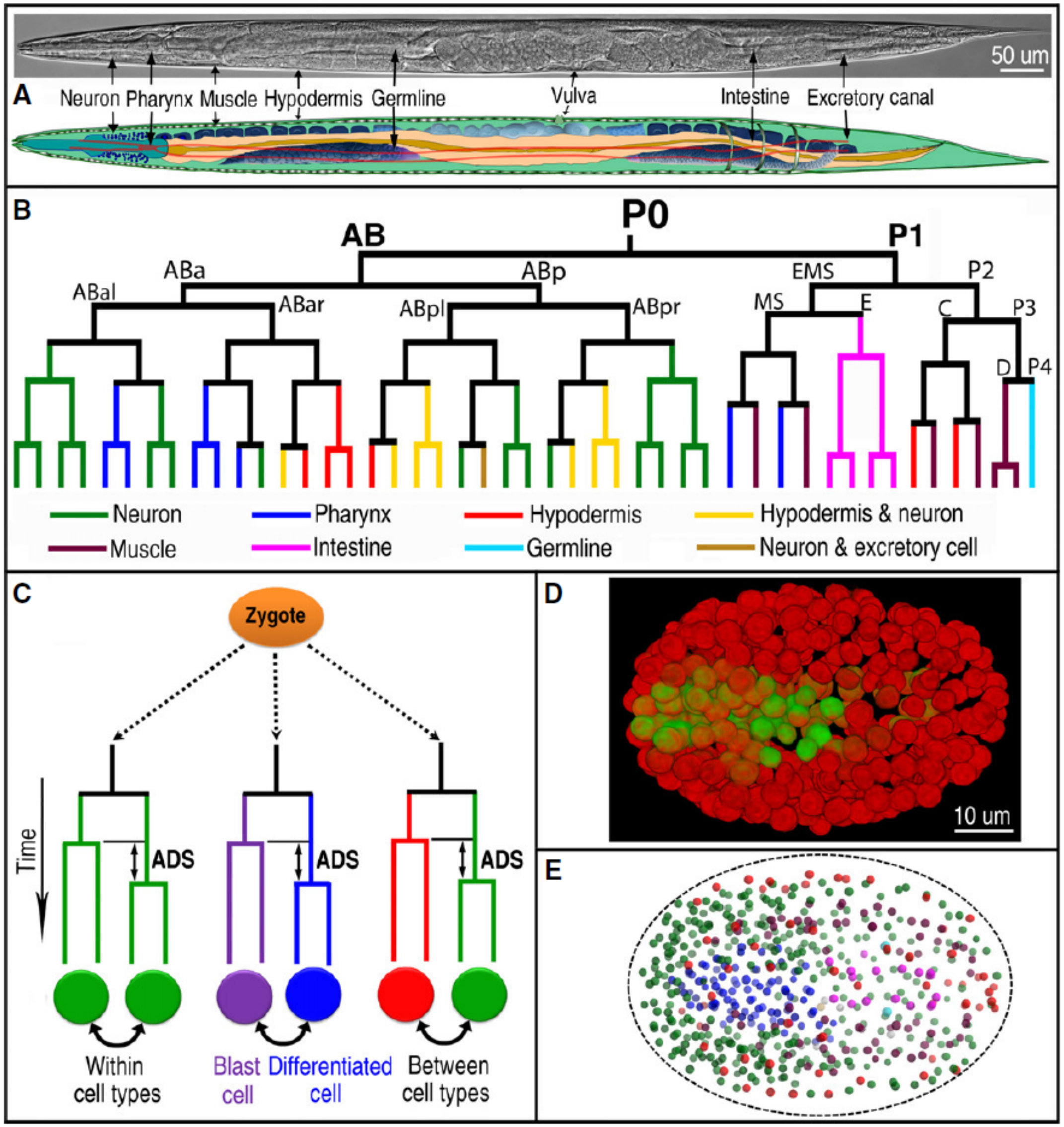






References: L Chen, Z Zhao, and H Yan, *IEEE J. Selected Topics in Signal Processing*, 10(1):185-192, 2016.
L Chen, Z Zhao, and H Yan, "Method for tracking an object in an image sequence," *United States Patent* 10,255,692, 2019.

C. Elegans Cell Division Lineage Tree



Tensor data:
- 1219 genes,
- 8 founder cells (AB branches),
- 14 descendants

References:

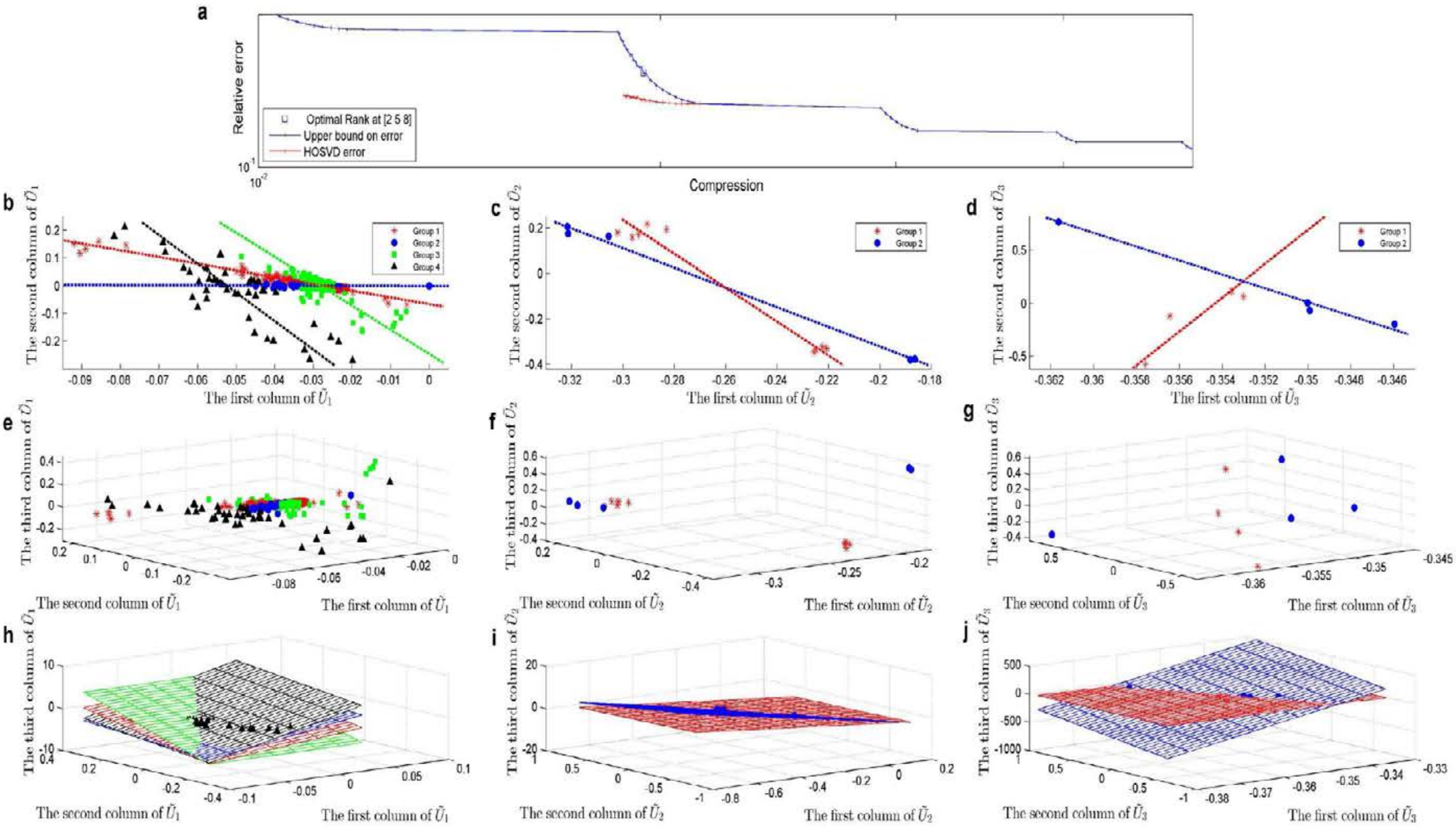
XT Huang, L Chen, H Chim, L Chan, Z Zhao, and H Yan, *BioMedical Engineering OnLine*, 12 (Suppl 1):S1, 2013.

VW S Ho, MK Wong, X An1, D Guan, J Shao, HCK Ng, X Ren, K He, J Liao, Y Ang, L Chen, X Huang, B Yan, Y Xia, LLH Chan, KL Chow, H Yan, and Z Zhao, *Molecular Systems Biology*,11:814, 2015.

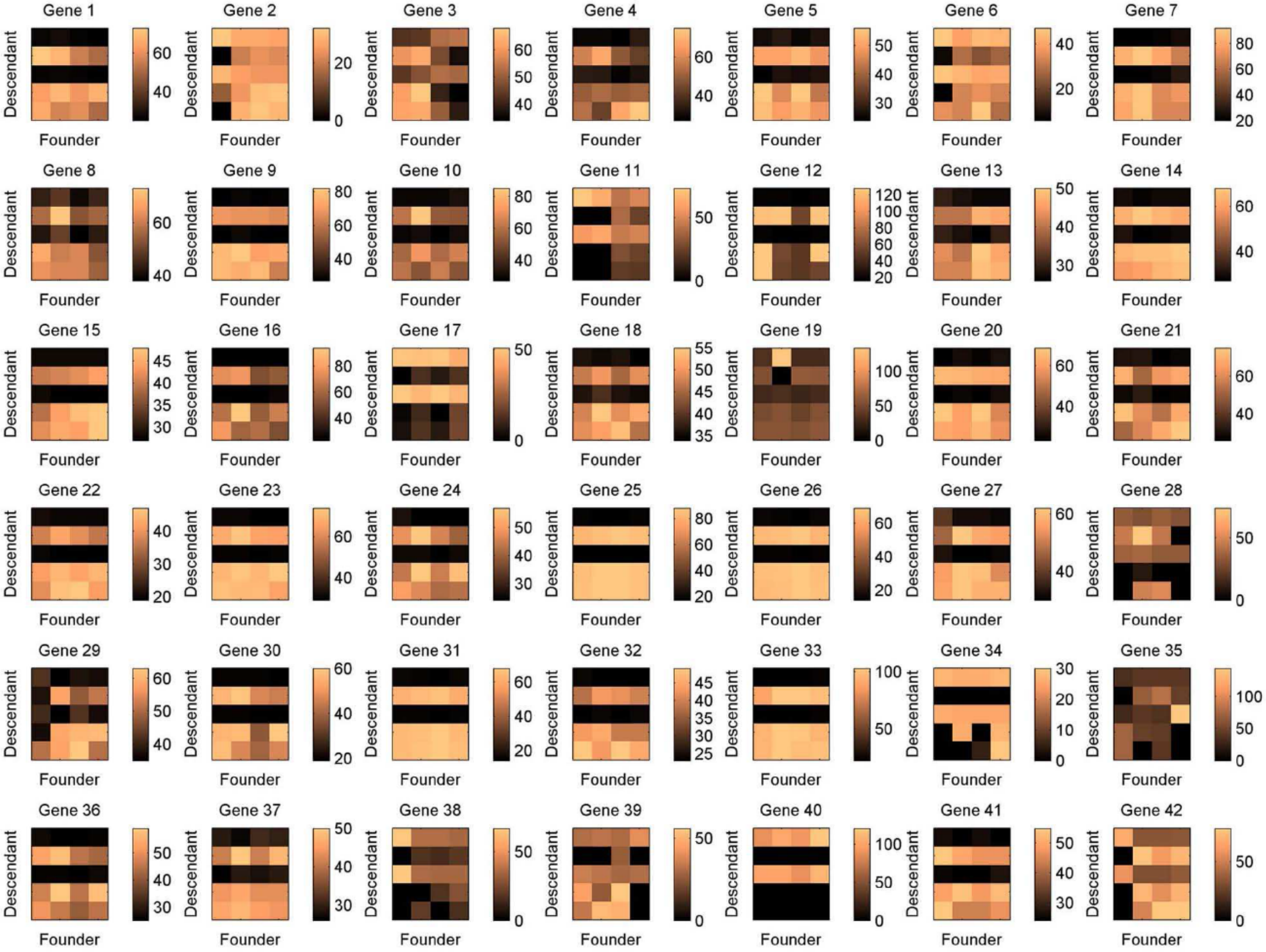
XT Huang, Y Zhu, LLH Chan, Z Zhao, and H Yan, *Molecular BioSystems*, 12:85-92, 2016.

J Cao, G Guan, VWS Ho, MK Wong, LY Chan, C Tang, ZY Zhao, and H Yan, *Nature Communications*, 11:6254:1-14, 2020.

Analysis of C. Elegans Data



Example of Co-cluster in C. Elegans Data



Feature Modes in C. Elegans Data

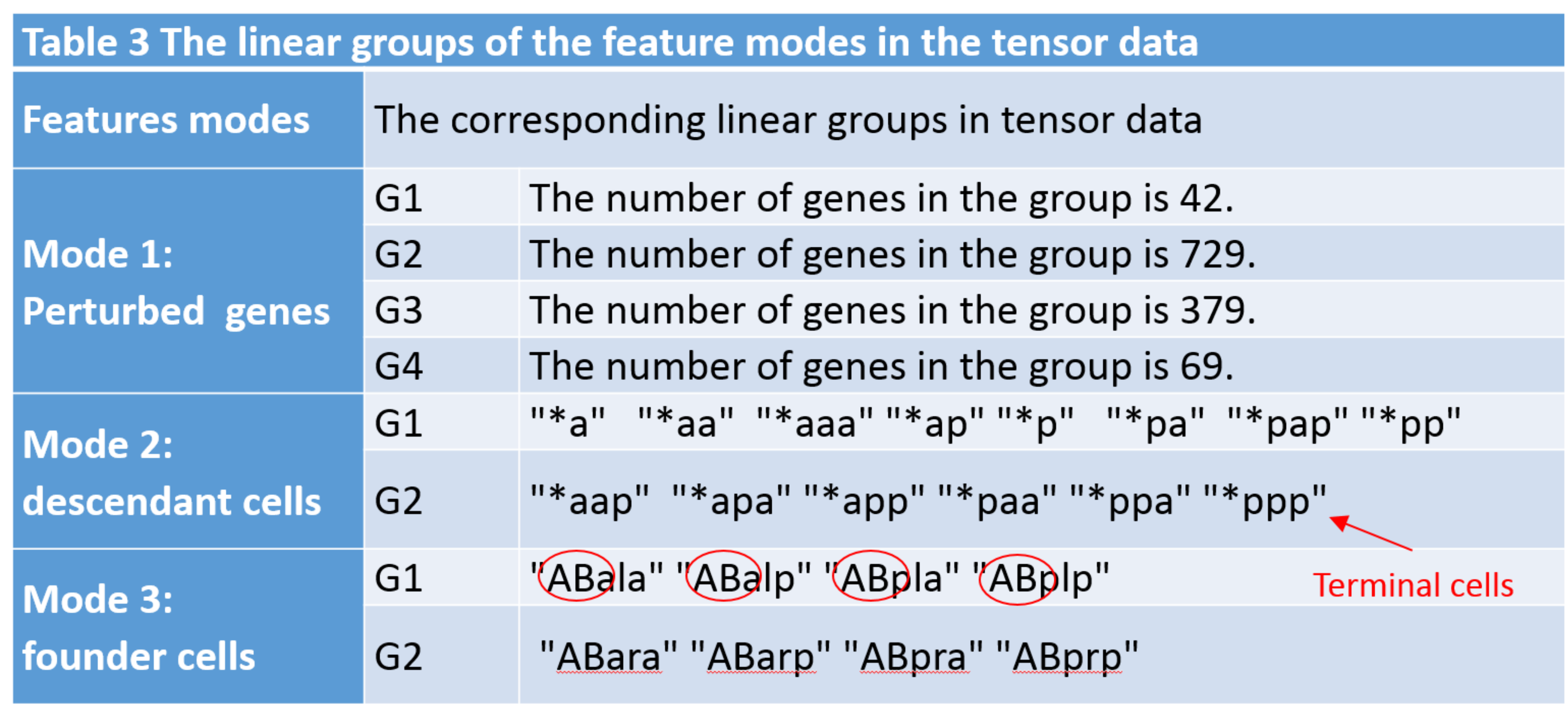


Table 3 The linear groups of the feature modes in the tensor data

Features modes	The corresponding linear groups in tensor data	
Mode 1: Perturbed genes	G1	The number of genes in the group is 42.
	G2	The number of genes in the group is 729.
	G3	The number of genes in the group is 379.
	G4	The number of genes in the group is 69.
Mode 2: descendant cells	G1	"*a" "*aa" "*aaa" "*ap" "*p" "*pa" "*pap" "*pp"
	G2	"*aap" "*apa" "*app" "*paa" "*ppa" "*ppp"
Mode 3: founder cells	G1	"ABala" "ABalp" "ABpla" "ABplp"
	G2	"ABara" "ABarp" "ABpra" "ABprp"

Terminal cells

GO / Pathway Analysis of C. Elegans Data

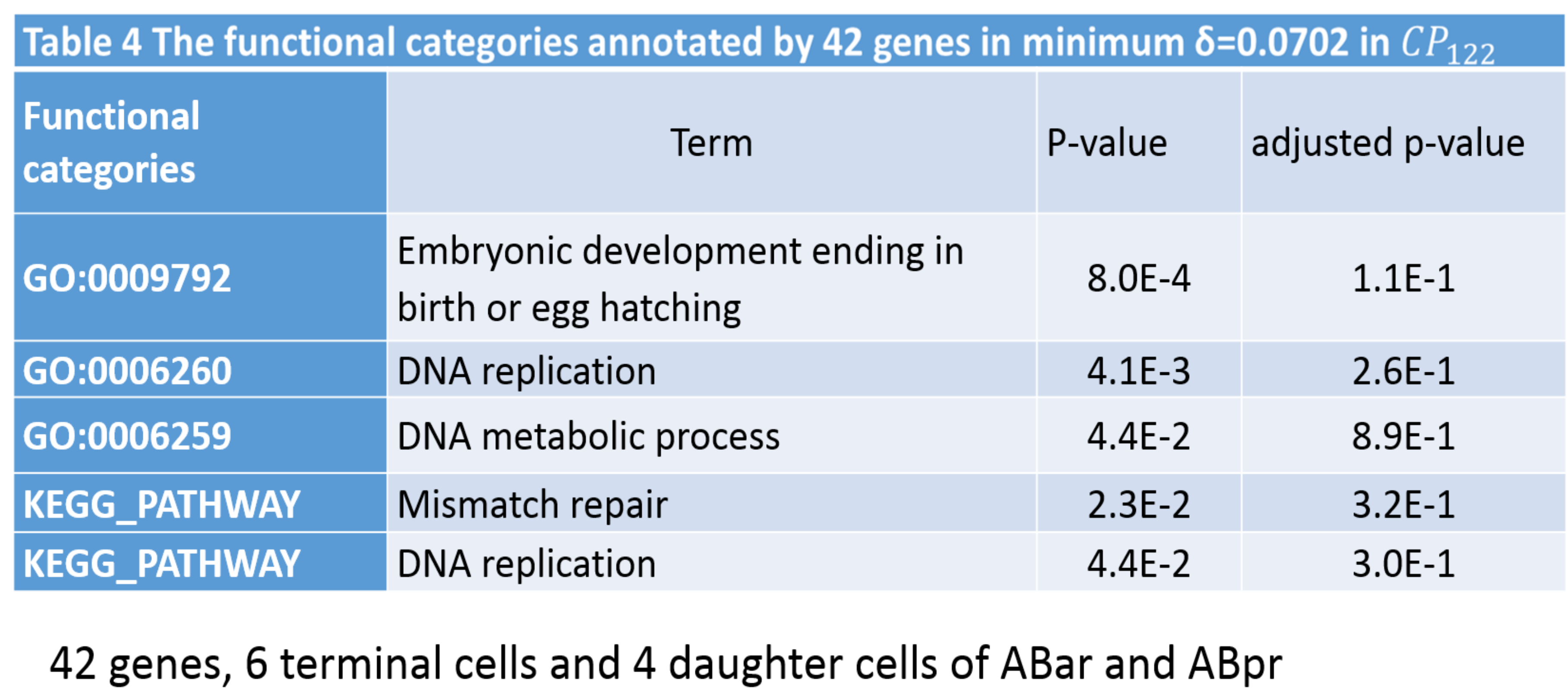


Table 4 The functional categories annotated by 42 genes in minimum δ=0.0702 in CP_{122}

Functional categories	Term	P-value	adjusted p-value
GO:0009792	Embryonic development ending in birth or egg hatching	8.0E-4	1.1E-1
GO:0006260	DNA replication	4.1E-3	2.6E-1
GO:0006259	DNA metabolic process	4.4E-2	8.9E-1
KEGG_PATHWAY	Mismatch repair	2.3E-2	3.2E-1
KEGG_PATHWAY	DNA replication	4.4E-2	3.0E-1

42 genes, 6 terminal cells and 4 daughter cells of ABar and ABpr

Protein Torsion Angles

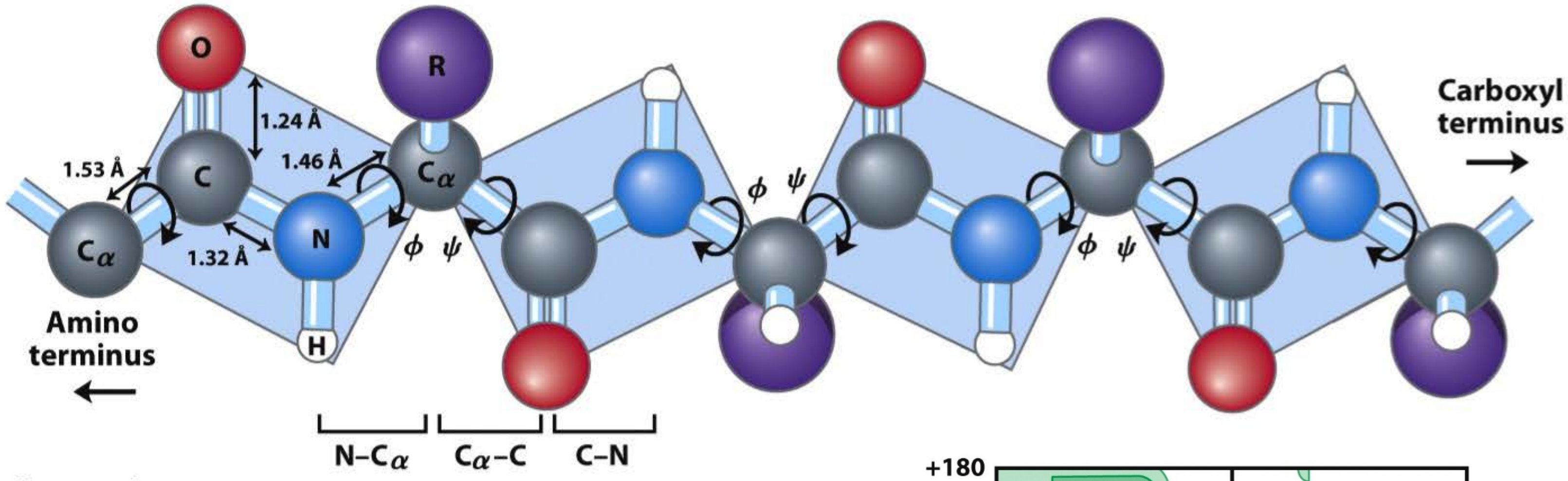


Figure 4-2b
Lehninger Principles of Biochemistry, Fifth Edition

Phi / Psi space:
N residues and 1^{o} sampling interval
360^{2N} possibilities
How to do the optimization: Levinthal's paradox

Reference:
https://www.studyblue.com/notes/note/n/chapter-4-1-biochemistry/deck/2294117
http://oregonstate.edu/instruct/bb450/450material/lecture/proteinstructureIoutline.html

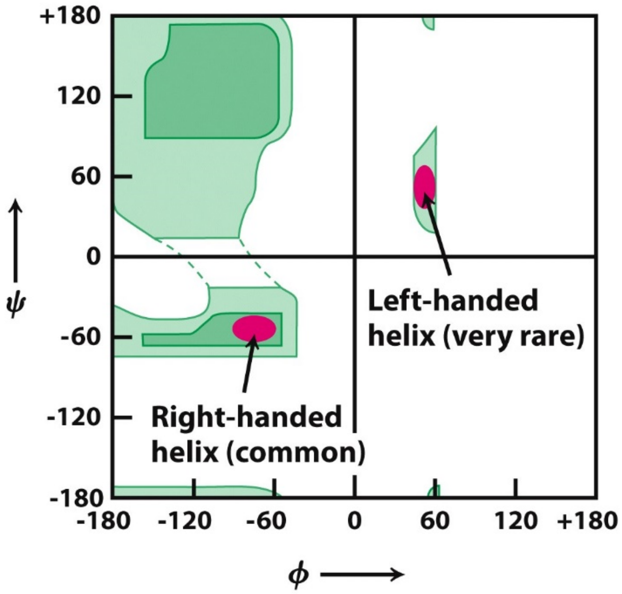
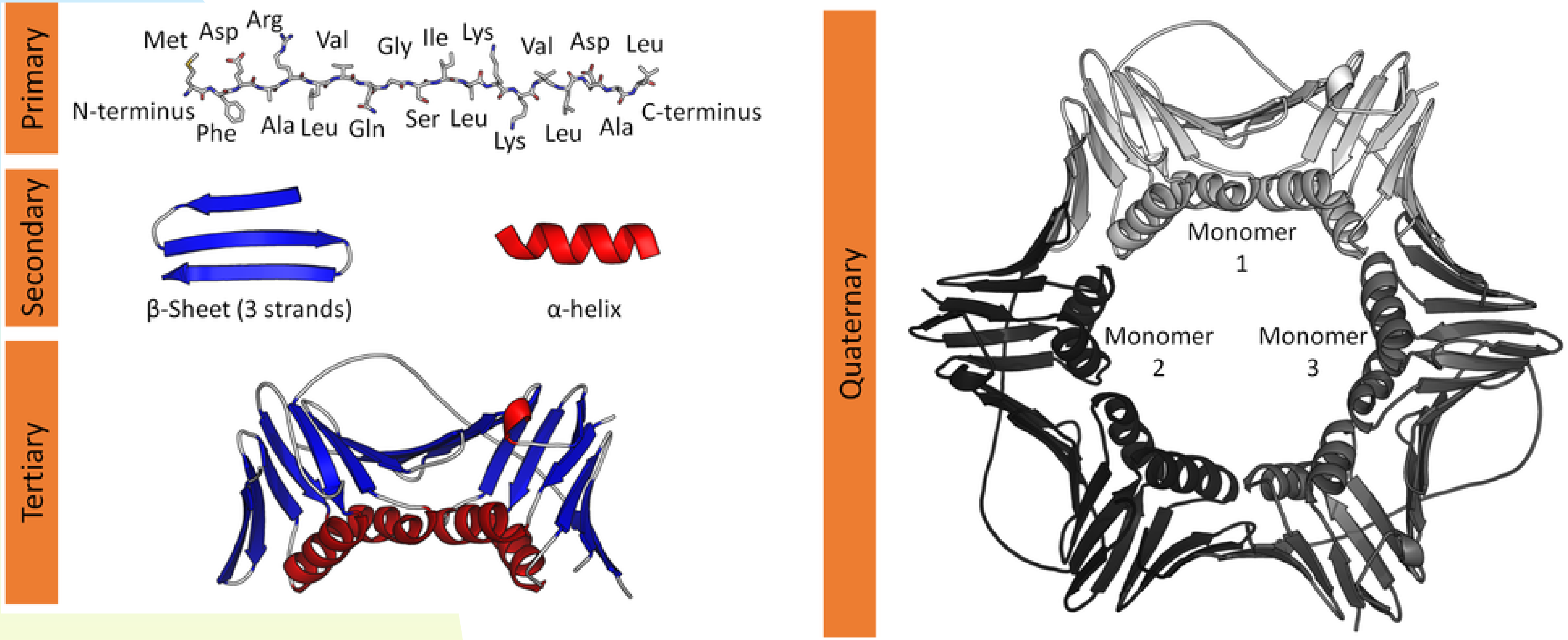


Figure 2.26
Biochemistry, Seventh Edition

Protein Structure

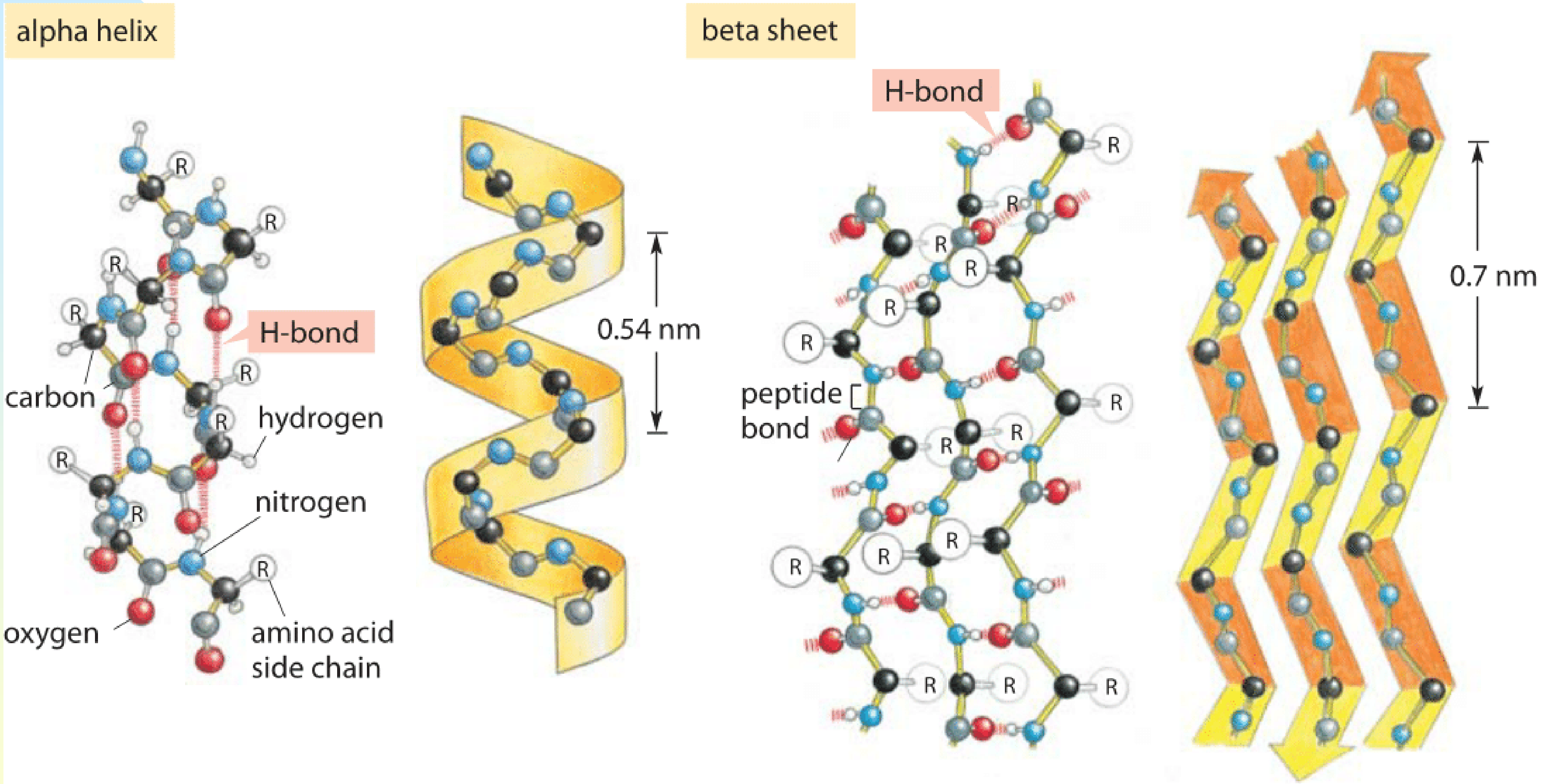


Reference: https://en.wikipedia.org/wiki/Protein_structure

Protein 3D structure prediction: "Holy Grail" problem

Related problems: Protein / DNA / RNA / ligand interactions

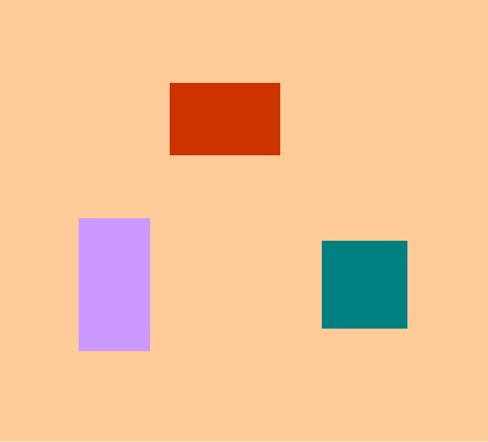
Alpha-Helix and Beta-Sheet



Reference: http://book.bionumbers.org/what-is-the-energy-of-a-hydrogen-bond/

Protein Secondary Structure Prediction

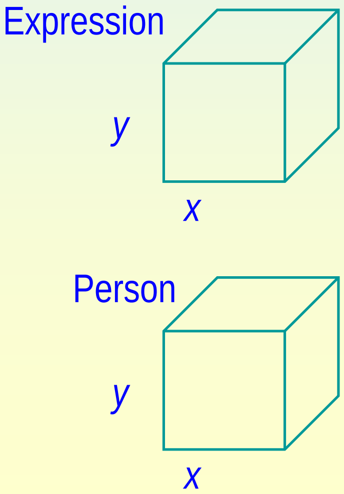
> 5000 features:
amino acids, neighbours, hydrogen bonds
15,310 α-helices, 20,847 β-strands
Co-clustering for feature selection
Selected features used for classification



Test	Three largest co-clusters (Row x Column)	Testing accuracy (1 co-cluster)	Testing accuracy (2 co-clusters)	Testing accuracy (3 co-clusters)
1	4284 x 10, 1696 x 9, 1800 x 7	0.8426	0.9630	0.9868
2	4270 x 10, 1706 x 9, 1379 x 8	0.8645	0.9706	1.0000
3	4314 x 10, 2214 x 8, 1358 x 8	0.8587	0.9447	0.9664
4	4314 x 10, 1772 x 9, 1638 x 8	0.8643	0.9640	0.9900
5	4420 x 10, 1733 x 9, 1454 x 8	0.8600	0.9633	0.9630

Reference: L Ma, DD Wang, X Liu, B Zou, and H Yan, *Current Bioinformatics*, 12(3):213-224, 2017.

Human Facial Expressions



Even higher dimensionality

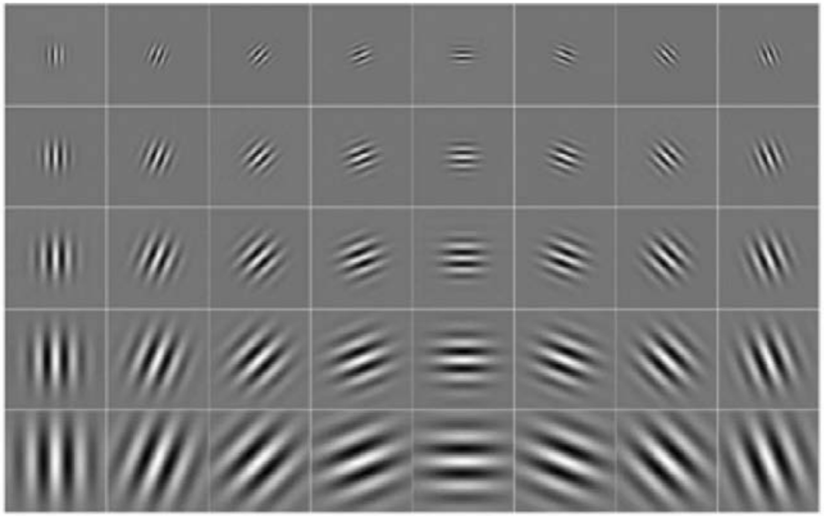
- Picture x Expression x Person x Age
- Picture x Expression x Directional wavelets

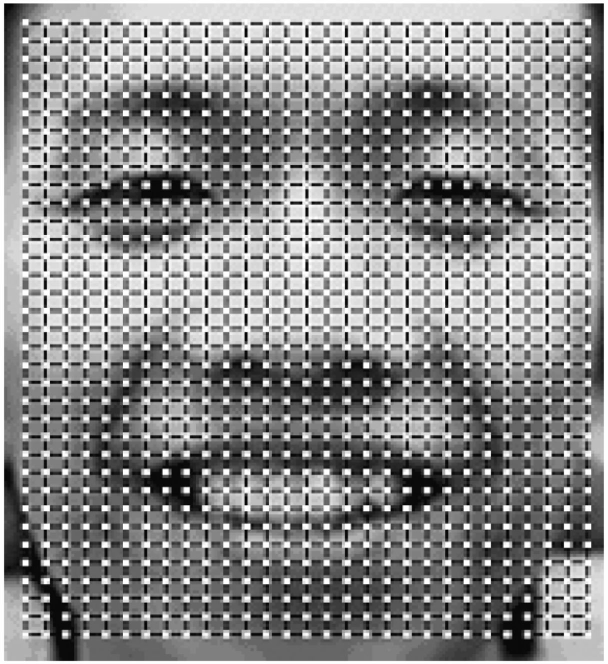
Gabor Wavelet Features

$$G_{\vec{k}}(\vec{r}\,) = G_{\vec{k},+}(\vec{r}\,) + iG_{\vec{k},-}(\vec{r}\,)$$

$$G_{\vec{k},+}(\vec{r}\,) = \frac{k^2}{\delta^2}\exp\left(\frac{k^2\|r - r_o\|^{\ 2}}{-2\delta^2}\right)\cos[\vec{k}(\vec{r} - \vec{r_o}\,)]$$

$$G_{\vec{k},-}(\vec{r}\,) = \frac{k^2}{\delta^2}\exp\left(\frac{k^2\|r - r_o\|^{\ 2}}{-2\delta^2}\right)\sin[\vec{k}(\vec{r} - \vec{r_o}\,)]$$



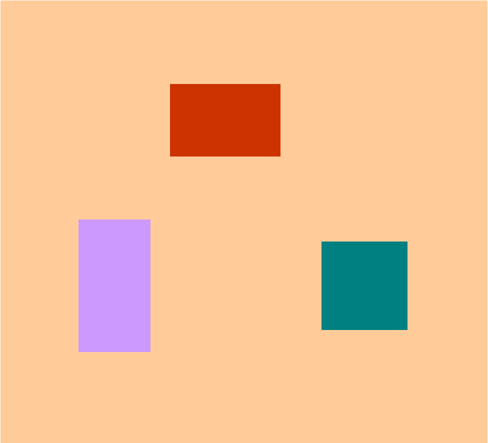


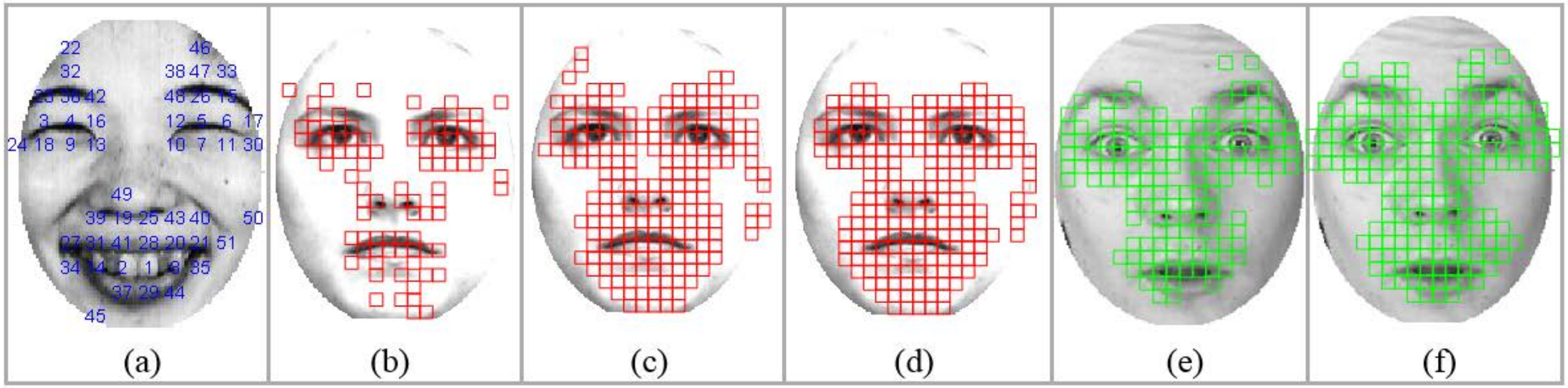
Reference: A Amin, H Yan, *Int'l J. Pattern Recognition and Artificial Intelligence*, 23(3): 401–431, 2009.

Feature Selection based on Co-clustering

537 samples, 19200 features
Co-clustering for feature selection
Selected features used for classification

All 19200 features used → 89.23%
3891 (20%) features retained → 96.14%
191 (0.9%) features retained → 90.23%




(a) (b) (c) (d) (e) (f)

Reference: S Khan, L Chen, and H Yan, *IEEE T Affective Computing,* 11(2):348-360, 2020.

Summary

- Coherent patterns may exist in multidimensional data
- Coherent patterns can be represented as low-rank matrices or tensors
- Coherent patterns can be detected in singular vector spaces
- Coherent patterns correspond to natural groups in multidimensional data
- Co-clustering can be used to analyze other “big data”

End of Presentation

E-mail: h.yan@cityu.edu.hk

Websites: http://www.ee.cityu.edu.hk/~hpyan

https://www.innocimda.com

Thank you

for

attending the presentation